Cracking the Code of the 2018 Democratic Primary Elections

Sara Tetsu, Aileen Wu, Nikki Iyer, and Carlos Gonzalez

December 11, 2023

https://github.com/tetsu2001/102-project

Introduction

Data Overview

The primary dataset we are using for this project is FiveThirtyEight's 2018 Primary Candidates dataset, which contains two separate comma-separated value files on the Democratic and Republican primary candidates for House, Senate, and gubernatorial races in districts where an incumbent was not a candidate in the election (FiveThirtyEight, 2022). FiveThirtyEight, Ballotpedia, and ABC News generated the data from campaign websites, news reports, state election websites, and official campaign filings (Conroy et al., 2018). In addition, we merged the FiveThirtyEight data with datasets from the Federal Election Commission (FEC) and The Brookings Institution because the original dataset does not contain information on campaign funding, candidate gender, and candidate beliefs, which we believe are important factors to consider in our analysis (Federal Election Commission, 2018; Kamarck & Podkul, 2018). The data can be viewed as a census since it includes every race that satisfies these conditions. Because election candidate data is assumed to be public information, the participants, or candidates, should be aware of its existence and differential privacy is not applicable.

Each observation in the FiveThirtyEight dataset represents one candidate in a regular or special House, Senate, or gubernatorial primary race. Data on each race are spread across multiple rows, and the same candidate may appear more than once in the dataset if they participated in both a regular and special election. Similarly, the rows in the FEC and Brookings datasets each represent one candidate. Because we want to study the effect of each candidate's attributes on their election outcome, this granularity is sufficient.

The only group excluded from this data is candidates in races where the incumbent sought reelection. Measurement error is not a significant concern, but during our data exploration, we found inconsistencies between the FiveThirtyEight and Brookings datasets for common features, such as military service. When this occurred, we defaulted to the Brookings data as it appeared to be more reliable. Convenience sampling is applicable here.

Candidates varied in the amount of information publicly available about them, meaning the amount of information we had on both winners and losers, varied person to person. Notable examples are a candidate's marital status, their position on climate change, or how various interest groups would have felt about them, had they weighed in on those races. Where feasible, reliable secondary sources are used to address these missing values, but it was also recognized that a lack of information may reflect a candidate's decision to not publicly disclose information, that is, no information *is* in itself, information. We used domain knowledge to decide whether a column's null values were systematic enough to be one-hot-encoded as their own column, or if they were infrequent enough that they could be treated as a "0" by default (e.g. if a candidate

was missing information on if they were an Obama alum or not, we interpreted the missing information as meaning they were not).

Our data cleaning and pre-processing one-hot encoding categorical variables where appropriate and filtering out special races and races where the candidate ran unopposed. We also restricted our analysis to House of Representatives races due to differences between House, Senate, and gubernatorial elections. Lastly, we used regex and fuzzy matching to merge the three datasets based on candidate name and manually checked for correctness. No FiveThirtyEight rows were lost in this process, and our final row count is 632.

Research Questions

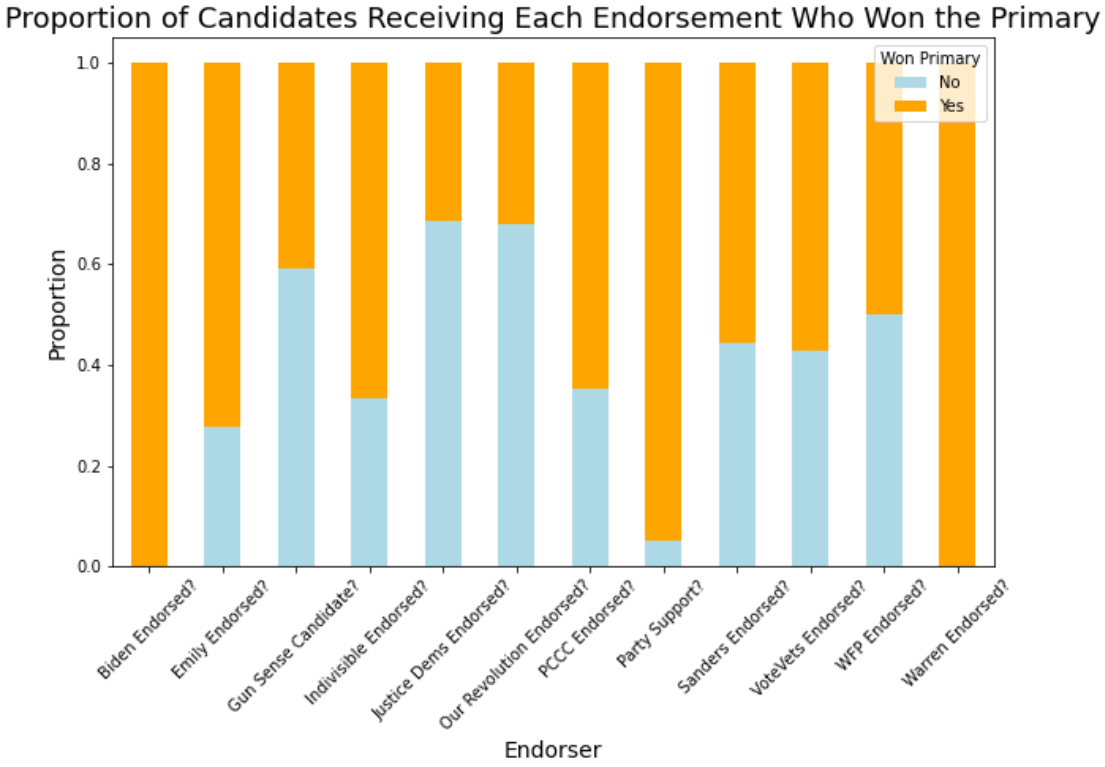
Using GLMs and nonparametric methods, we aim to discern key factors influencing Democratic primary wins in 2018, striving for both high accuracy and clarity. This analysis could clarify prevailing issues and predict future candidate success, aiding in diversifying Congress representation (Schaeffer, 2023).

Our first objective is, broadly speaking, to understand the true relationship, which we will denote as F(X), between candidates and their election outcomes. Because the goal of estimating F is inference, we chose to fit a logistic regression model of the form $Y = \text{sigmoid}(X^T\beta)$. A logistic regression model, like all parametric approaches, bring with it the possibility that the functional form used to estimate F is very different from the true F. In this case, in fitting a logistic regression model, we are making an implicit assumption of linear interactions between candidate information, X, and election outcome, Y. The resulting model is ideal for our objective of drawing inferential conclusions because the feature parameters can be interpreted in a meaningful way. If feature *x* increases by a certain amount *a*, then the log-odds of Y=1 increases by $\beta*a$.

On the other hand, we aim to use causal inference to answer our second question, which is whether or not the endorsement by certain political organizations, such as political action committees (PACs), individually have a causal effect on the success of candidates in the 2018 Democratic primary elections. In a sense, this second question is an extension of the first one. By analyzing the causal effect of various types of endorsements—from progressive PACs to establishment groups—we can learn about whether certain types of endorsements are more influential than others. While existing literature has explored the association between endorsements and election outcomes in a two-party general election, causal inference would reinforce those findings by isolating endorsement effects and provide novel insights into the effectiveness of endorsements in the unique conditions of a nationwide primary election, where candidates of the same party are vying for victories (Boudreau, 2016). As with any causal inference model, the results of this analysis are contingent upon how exhaustive our confounders are. If we fail to block all backdoor paths, mistake a collider for a confounder, or simply lack crucial data on confounders, the model will not accurately estimate the true causal effect of an endorsement on winning elections.

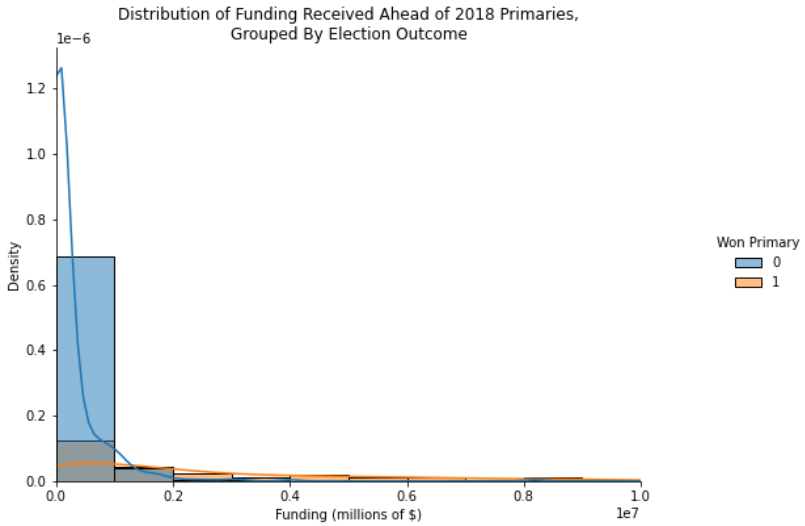
Exploratory Data Analysis

Figure 1. Stacked barplot showing the proportion of candidates who won the primary, grouped by endorsement. The granularity is one endorsement given, so the same candidate can appear more than once if they received multiple endorsements.



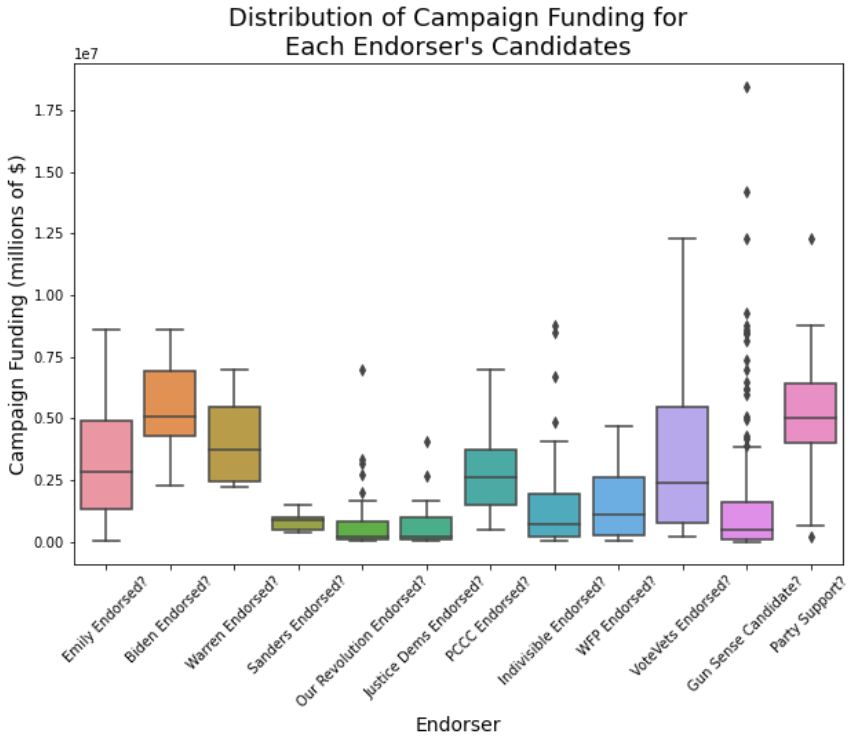
Among the endorsers, Biden, Warren, and the Democratic Party had the greatest proportion of winning candidates. However, certain endorsers could have chosen their candidates more strategically. Since Biden, Sanders, and Warren each gave between five to ten endorsements, this may have especially been the case for them. Among the PACs, Justice Dems and Our Revolution had the lowest proportion of winning endorsements, whereas Emily's List, Indivisible, and PCCC saw the greatest success. Interestingly, the two worst-performing endorsements are both progressive PACs. This suggests that not all endorsements have the same influence, so we should not expect them to have the same causal effect.

Figure 2. Overlaid histogram of financial contributions received by each campaign in 2018, grouped by election outcome. Capped at $1 million for visualization purposes only.



More candidates who won their election had more funding, indicated by how the orange density line rises above the blue line as funding increases.

Figure 3. Boxplot showing the distribution of campaign funding received by candidates per endorser.

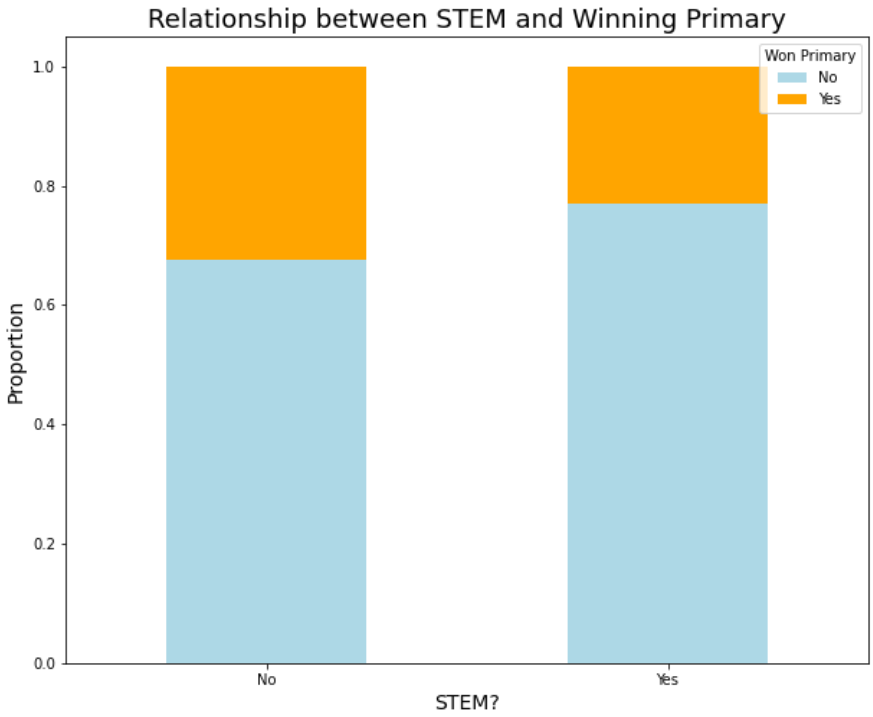


Based on figures 2 and 3, campaign funding should be a candidate when considering what confounding variables to include in causal inference. Figure 3 shows how many candidates

endorsed by progressive PACs had less funding than candidates associated with establishment organizations.

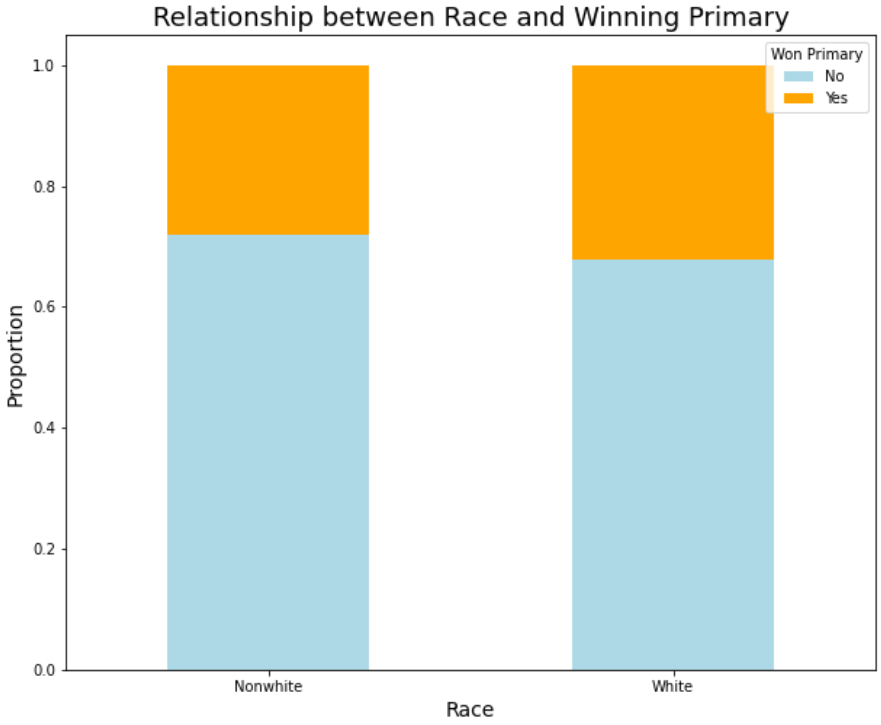
The majority of demographic features showed nothing remarkable. However, we included visualizations for variables that did appear to contain some association with success.

Figure 4. Proportions of STEM and non-STEM candidates who won their race.



More candidates with non-STEM backgrounds won their primary compared to candidates with STEM backgrounds. This could suggest that voters do not view having a STEM background as a particularly desirable or important trait, or it could imply that candidates with a STEM background have less campaign experience or connections than, for example, candidates from a political science background who have had more exposure to the political world.

Figure 5. Proportions of white and non-white candidates who won their race.



White candidates fared better than non-white candidates. Out of the candidates who were white, there was a slightly higher percentage of people who won the primary. Because there appears to be a difference, we want to incorporate race into our models to accurately capture the relationships between the variables.

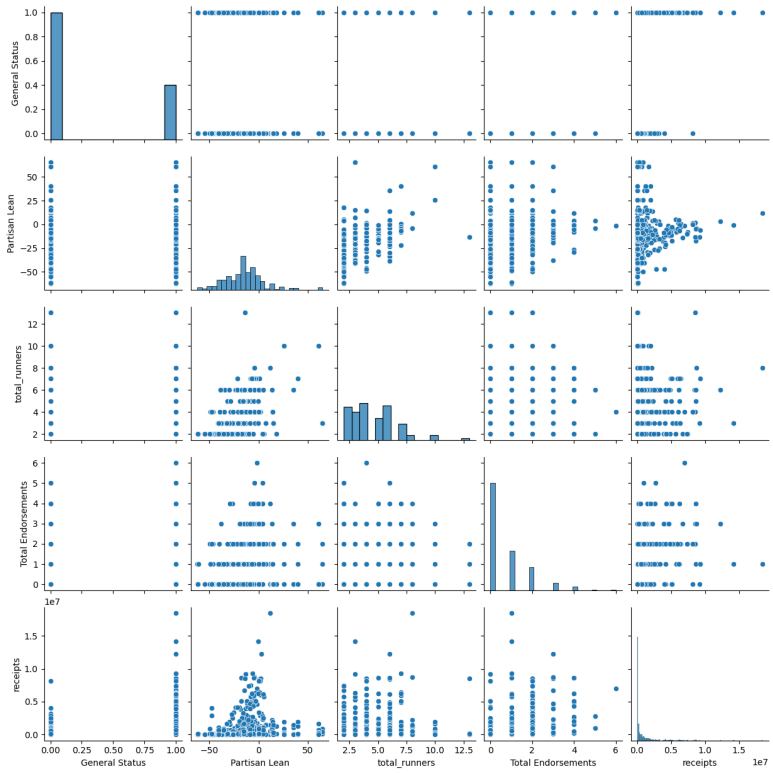
Methods

Prediction with GLMs and Nonparametric Methods

The dataset includes 48 predictors: 44 categorical and 4 numerical. The outcome, Y, is also a categorical variable, that takes on the value 1 if the candidate won the primary, and 0 if the candidate did not. To select a model, we first used a chi-squared test on the categorical predictors and Pearson correlation-coefficient test on the numerical. These help us answer the question, “Are any of the predictors in our dataset significantly associated with a candidate's outcome?" The null hypothesis, that all regressors are not significantly associated with outcome, is rejected if the p-value for at least one predictor in our model is below 0.05. By running these tests, we determined that only 24 of the predictors in our dataset had a statistically significant association with candidate outcome. These predictors and are plotted below:

Figure 6: Stacked bar plots of proportion of features with relation to winning the election

Figure 7: Pair plots of various different significant predictors and their relation to general status



We then drop all other predictors from our dataframe and one-hot-encoded the categorical variables of this reduced dataframe. To avoid multicollinearity, we dropped one category per one-hot-encoded variable. Rather than dropping any random category, we chose to drop the category with the least interpretability, such as “Marital Status: No Information” and “Climate Change: Candidate provides unclear position” In doing this, we were able to rid our dataframe of multicollinearity while maintaining model interpretability. While our model did not have exact linear dependence, several of the columns of our dataframe were highly correlated with one another. To address this, we wrote a function to calculate the VIF, or variance inflation factor of the columns in the dataframe, remove the variable with the highest VIF from the dataset, and then calculate the VIF on the data frame again, dropping the column with the highest VIF each time. We continued this process iteratively until the VIF was below 2. This allowed us to reduce our dataset to 19 predictor columns. A model with 19 predictors means there are 2^{19} possible logistic regression models. Rather than trying all of them manually (a computationally infeasible and luckily unnecessary task) we used a forward-selection function to iteratively add predictors to the model, and selected the model with the lowest AIC to balance model fit and complexity.

Figure 8: Summary statistics of the generalized linear model

"""
Generalized Linear Model Regression Results

Dep. Variable:	General Status	No. Observations:	632
Model:	GLM	Df Residuals:	616
Model Family:	Binomial	Df Model:	15
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-232.51
Date:	Mon, 11 Dec 2023	Deviance:	465.03
Time:	19:04:40	Pearson chi2:	4.33e+03
No. Iterations:	21	Pseudo R-squ. (CS):	0.3855
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-3.9162	0.367	-10.676	0.000	-4.635	-3.197
Position.on.Defense.Spending_Candidate supports a reduction in military spending	-1.0048	0.653	-1.540	0.124	-2.284	0.274
Indivisible Endorsed?_1.0	0.3799	0.482	0.788	0.431	-0.565	1.325
Emily Endorsed?_1.0	0.3898	0.645	0.605	0.545	-0.874	1.653
Party.Category_Establishment Democrat	0.4523	0.248	1.824	0.068	-0.034	0.938
Our Revolution Endorsed?_1.0	1.0080	0.363	2.779	0.005	0.297	1.719
VoteVets Endorsed?_1.0	-1.1500	0.860	-1.337	0.181	-2.835	0.536
Female_1	1.3375	0.251	5.332	0.000	0.846	1.829
STEM?_1	-0.5910	0.324	-1.821	0.069	-1.227	0.045
PCCC Endorsed?_1.0	1.4787	1.232	1.201	0.230	-0.935	3.893
Partisan Lean	-0.0579	0.008	-7.266	0.000	-0.073	-0.042
Position.on.Federal.K.12.Education.Policy_Candidate supports federal proposals for major education reform (including common core)	0.1950	0.246	0.794	0.427	-0.287	0.677
Education_J.D.	0.5971	0.306	1.949	0.051	-0.003	1.198
receipts	1.407e-06	2.05e-07	6.881	0.000	1.01e-06	1.81e-06
Biden Endorsed?_1.0	20.3123	1.28e+04	0.002	0.999	-2.51e+04	2.52e+04
Gun Sense Candidate?_1.0	0.6138	0.272	2.254	0.024	0.080	1.148

"""

The model with the lowest AIC used 10 predictors. To assess how well this model actually fits our data, we ran a logistic regression model using 5-cross validation.

We built the non-parametric models employing the same features, in order to extend our conclusions to be able to predict the success of candidates in future midterm elections. We opt to use both decision trees and random forests because these methods do not make explicit assumptions about the functional form of F. Instead they seek an estimate of f that gets as close to the data points as possible without being too rough or wiggly. Another constraint when designing the nonparametric methods lies in our observations of the present data: 30% of the entries in our data contained candidates which did not advance past the general election, meaning that simply interpreting our nonparametric methods' performance via accuracy scores would be inideal. We implemented random over-sampling (ROS) and random under-sampling (RUS) to achieve class balance in the train and test sets. In addition to accuracy, we looked at confusion matrix metrics, such as precision, recall, F1 score, mean cross-validation scores, ROC curves, and AUC values to validate our findings.

Causal Inference

The treatment is receiving an endorsement from a specific, individual PAC ahead of the 2018 primary election, and the outcome is whether or not the House of Representatives candidate won their election ('Won Primary'). Because different endorsements likely had different effects and some politicians or organizations did not endorse enough candidates, we specifically examined seven separate treatments that had at least 90 non-NaN instances in the cleaned dataset: receiving support from the Democratic Party ('Party Support?'), being endorsed by Our Revolution ('Our Revolution Endorsed?'), being endorsed by Justice Dems ('Justice Dems Endorsed?'), being endorsed by Indivisible ('Indivisible Endorsed?'), being endorsed by Emily's List ('Emily Endorsed?'), being endorsed by VoteVets ('VoteVets Endorsed?'), and being a Gun Sense candidate ('Gun Sense Candidate?').

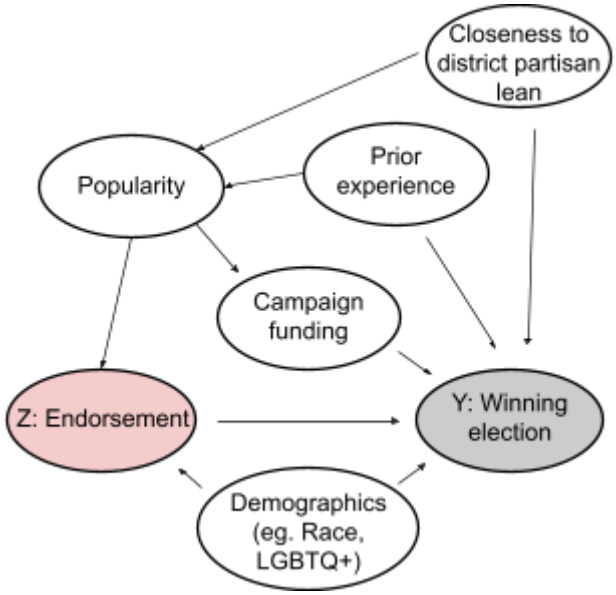
Numerous factors can influence a candidate's likelihood of winning, such as 'Previous.Electoral.Experience' indicating familiarity with the electorate and campaign expertise.

Popularity, often linked to the number of endorsements, could affect funding levels, as more recognized candidates tend to attract more donations. This funding can then enhance a candidate's visibility, further increasing their popularity. We also observed that a candidate's alignment with the partisan leanings of their district can influence both their endorsement prospects and election outcomes. Demographic factors like race, veteran status, and sexual orientation also play a role, impacting both a candidate's endorsements and their success rate. This results in a very complex web of features where factors such as experience, popularity, partisan alignment, funding, and demographics intertwine, all affect both endorsements and election results.

To adjust for confounders, we will incorporate inverse propensity weighting (IPW) with propensity scores calculated from logistic regression into our process of estimating the average treatment effect (ATE). Because rare observations that have very large or very small propensity scores may increase the variance of the estimator, we will also provide the trimmed IPW estimates by restricting to propensity scores between 0.1 and 0.9. Since the dataset contains relatively few observations, matching would be difficult and likely unfeasible.

There are no colliders in the dataset. Since the dataset only contains non-incumbent primary races, it is unlikely that the outcome, winning the primary election, could causally influence any of these variables.

Figure 9. A Causal DAG.

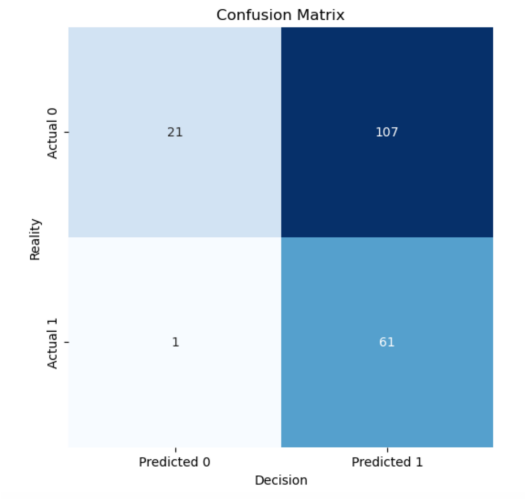


Results

Prediction with GLMs and Nonparametric Methods

To assess how well our forward selection model fit the data, we used 5-fold cross validation on the logistic model and used the results in the confusion matrix to calculate average precision, recall, and accuracy.

Figure 10: Confusion matrix of the GLM without Cross-Validation



The results are in the table below:

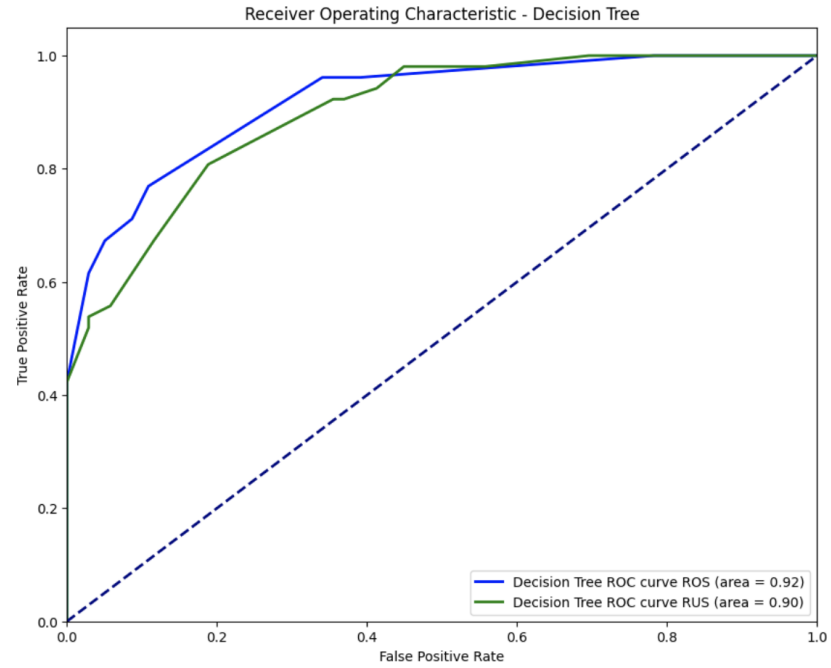
Average_precision	0.7517
Average recall	0.9737
Accuracy	0.4274
ROC_auc	0.8211
Prevalence	0.3006

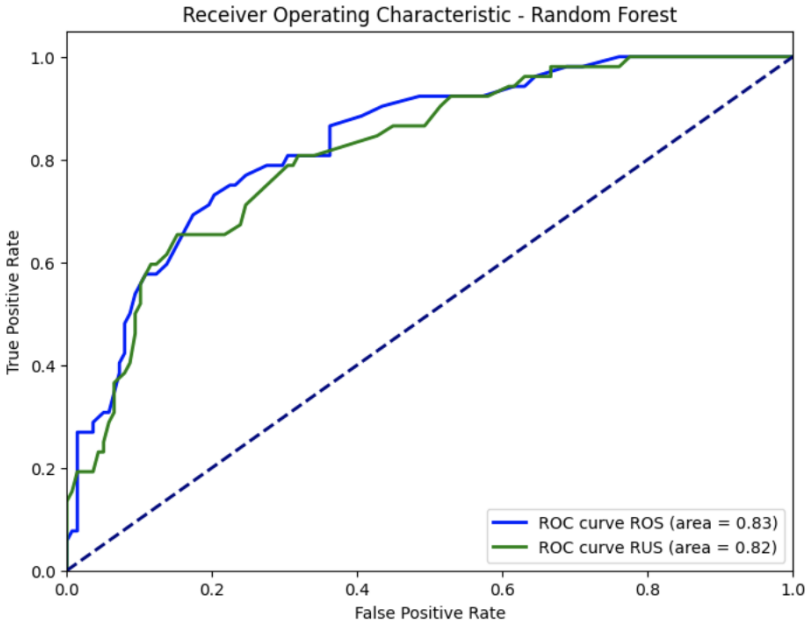
Our model had a high average recall, meaning that of all the winners, our model correctly identified 97% of them. The average precision of 0.75 means that of all the discoveries we made, approximately 25% of them were false discoveries. Our accuracy was 0.43, which is lower than the naive estimate. While disappointing, this makes sense: there are a lot of things that contribute to candidate success, and our dataset. However, we decided that accuracy was not a good metric to assess our model's fit because the naive estimate would give a 70% accuracy, but that does not mean the model would not be effective. Due to the dataset size and variance, we used average precision and recall as measures to validate the effectiveness of our model instead.

Of all of the features that we used on our model, we found that P(Y=1) depends on being female, being a A Gun Sense Candidate, monetary donations, having a STEM background, Partisan Lean of the district they ran to represent, and OurRevolution endorsements. These features demonstrated statistical significance with a p-value of less than 0.05. Being female-identifying is associated with an increase in the log odds of Y=1 by 1.2585 units.

We derived a decision tree model and applied ROS with an accuracy score of 72.63% with a maximum depth of five layers and a mean cross-validation score of 77.8%, whereas the decision tree model under RUS achieved an accuracy score of 73.68% and mean cross-validation score of 75.49%. The AUC remains the same for both ROS and RUS at 0.92 and 0.92, indicating consistent performance across sampling methods. The random forest model performs with an accuracy of 81% and a mean cross-validation score of 91% for ROS and an accuracy of 75% and a mean cross-validation score of 75% for RUS, with the AUC for the ROS and RUS to be 0.82 and 0.83, respectively. The similarity between the AUC and performance metrics for the two methods with ROS and RUS imply the model's ability to discriminate between the two target variable outcomes.

Figure 11. The ROC Curves of the Decision Tree & Random Forest Models under ROS and RUS





We also visualized the most important features used in making a decision for the decision tree.

Figure 12. Bar chart of the most significant features with decision tree over sampling

(NOTE: 5th entry on the bar plot reads: 'Position.on.Legalization.Decriminalization.of.Marijuana.Policy')

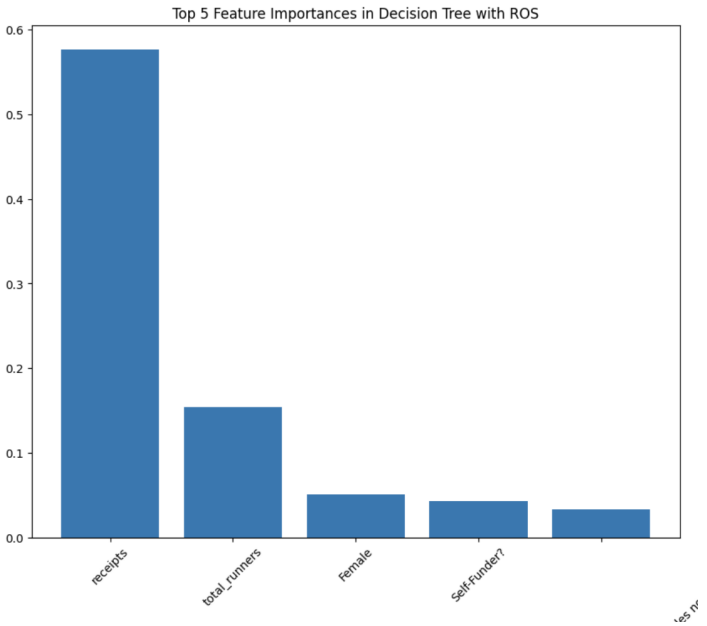
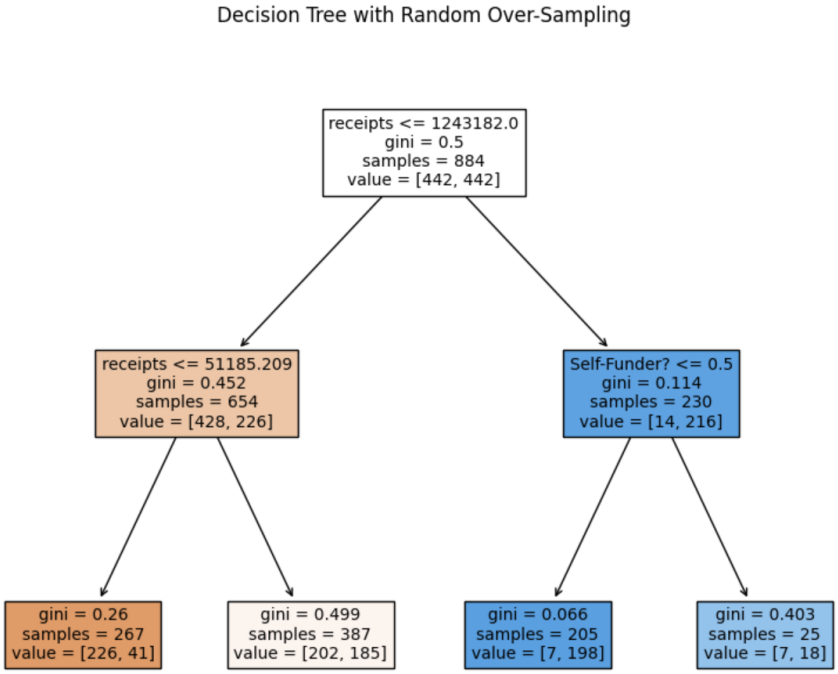


Figure 13. Decision Tree ROS Model with a maximum depth of 2



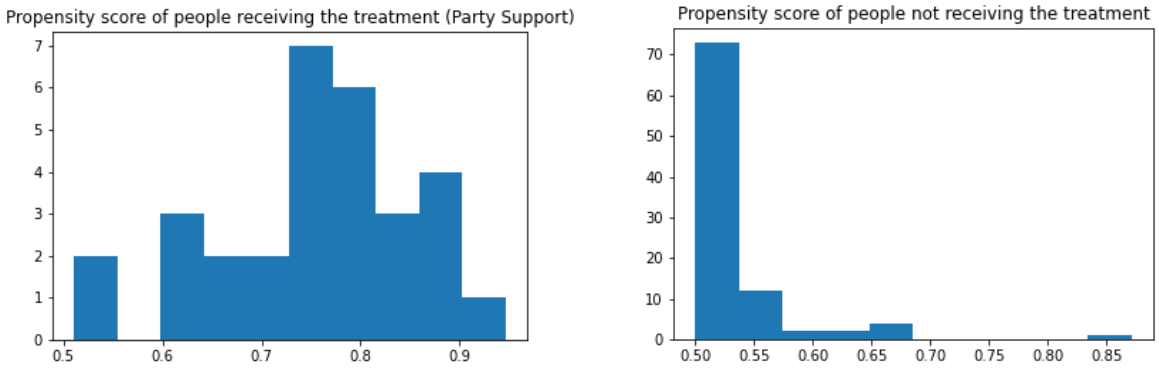
Causal Inference

Treatment 1: Receiving Democratic Party support

Without accounting for confounders, the naive estimate is 0.9014, which implies that receiving an endorsement from the Democratic Party increases the probability of winning by 90.14%. However, after applying inverse propensity weighting, the estimated average treatment effect became 0.2285, and the trimmed average treatment effect became 0.2218. Therefore, after controlling for confounders, an endorsement from the Democratic Party causes a 22.18%

increase in the probability of winning the primary election. The 95% confidence interval is [0.122, 0.337], indicating a statistically significant effect.

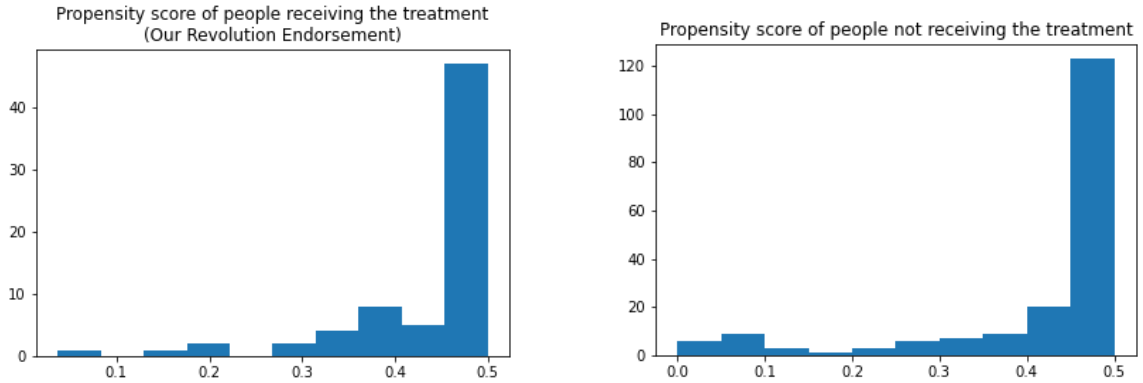
Figure 14. Histograms of propensity scores, grouped by treatment (Party Support)



Treatment 2: Receiving an Our Revolution endorsement

Without accounting for confounders, the naive estimate is 0.0897, which implies that receiving an endorsement from Our Revolution increases the probability of winning by 8.97%. After applying inverse propensity weighting, the estimated average treatment effect became 0.1046, and the trimmed average treatment effect became 0.0642. After controlling for confounders, an endorsement from Our Revolution causes a 6.42% increase in the probability of winning the primary election. However, the 95% confidence interval is [-0.038, 0.398], which contains 0. Thus, the effect is not statistically significant.

Figure 15. Histograms of propensity scores, grouped by treatment (Our Revolution Endorsed)

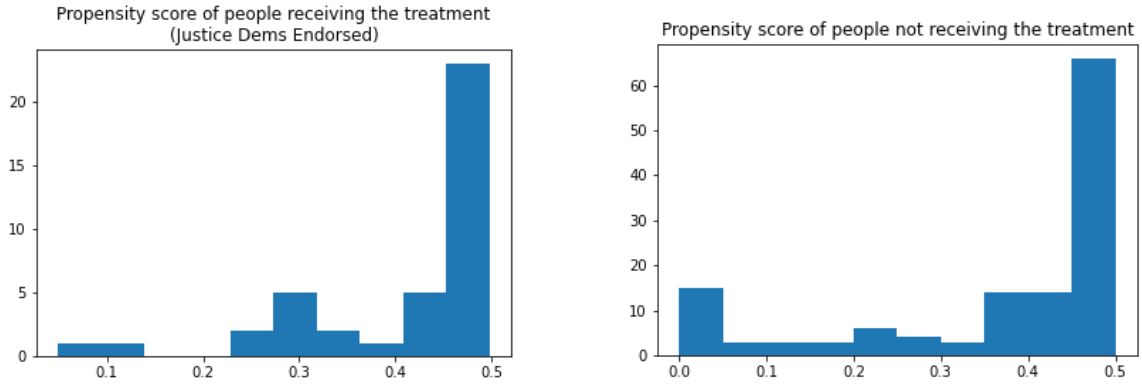


Treatment 3: Receiving a Justice Dems endorsement

Without accounting for confounders, the naive estimate is 0.0439, which implies that receiving an endorsement from the Justice Democrats increases the probability of winning by 4.39%. After applying inverse propensity weighting, the estimated average treatment effect became 0.1173, and the trimmed average treatment effect became 0.1123. After controlling for

confounders, an endorsement from the Justice Democrats causes an 11.23% increase in the probability of winning the primary election with a 95% confidence interval of [-0.055, 0.527]. Since this includes 0, the result is not statistically significant.

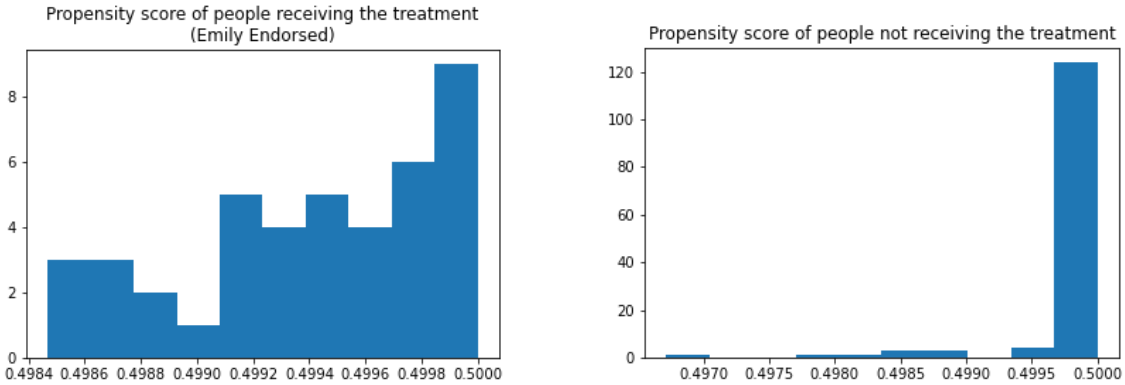
Figure 16. Histograms of propensity scores, grouped by treatment (Justice Dems Endorsed)



Treatment 4: Receiving an Emily's List endorsement

Without accounting for confounders, the naive estimate is 0.6102, which implies that receiving an endorsement from Emily's List increases the probability of winning by 61.02%. After applying inverse propensity weighting, the estimated average treatment effect and the trimmed estimate both became 0.2020. After controlling for confounders, an endorsement from Emily's List causes a 20.2% increase in the probability of winning the primary with a 95% confidence interval of [0.125, 0.277], which is statistically significant.

Figure 17. Histograms of propensity scores, grouped by treatment (Emily Endorsed)

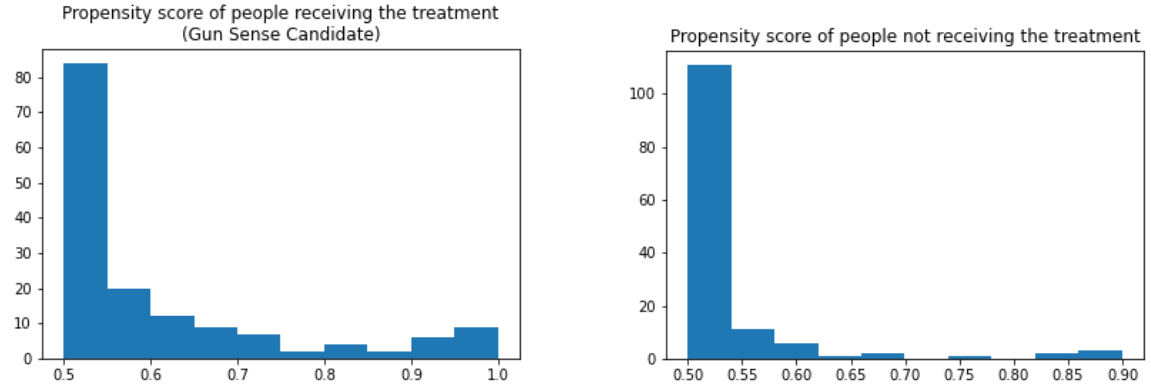


Treatment 5: Being a Gun Sense candidate

Without accounting for confounders, the naive estimate is 0.318, which implies that being a Gun Sense candidate increases the probability of winning by 31.8%. After applying inverse propensity weighting, the estimated average treatment effect became 0.1324, and the trimmed average treatment effect became 0.0827. After controlling for confounders, being a Gun Sense

candidate causes an 8.27% increase in the probability of winning the primary election with a 95% confidence interval of [-0.007, 0.214], which is not statistically significant.

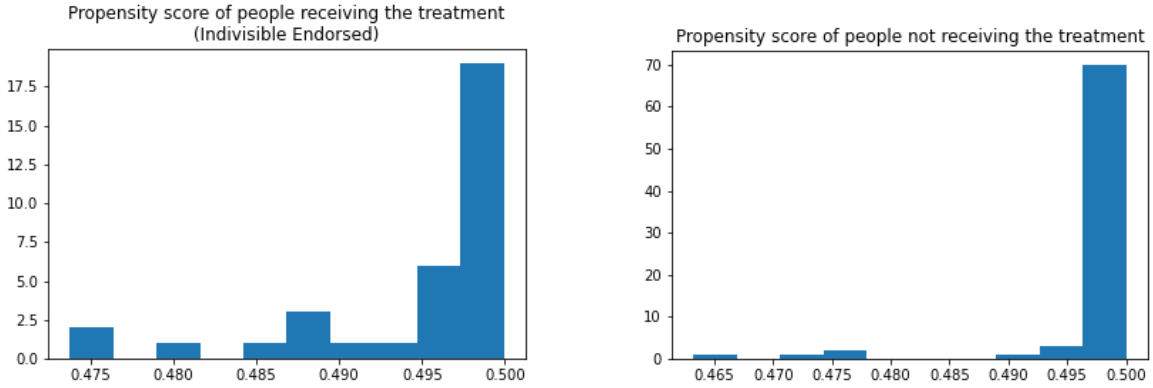
Figure 18. Histograms of propensity scores, grouped by treatment (Gun Sense Candidate)



Treatment 6: Receiving an Indivisible endorsement

Without accounting for confounders, the naive estimate is 0.46, which implies that being endorsed by Indivisible increases the probability of winning by 46%. After applying inverse propensity weighting, the estimated average treatment effect and the trimmed estimate both became 0.1885. After controlling for confounders, an Indivisible endorsement causes an 18.85% increase in the probability of winning the primary election with a 95% confidence interval of [0.045, 0.331], which is statistically significant.

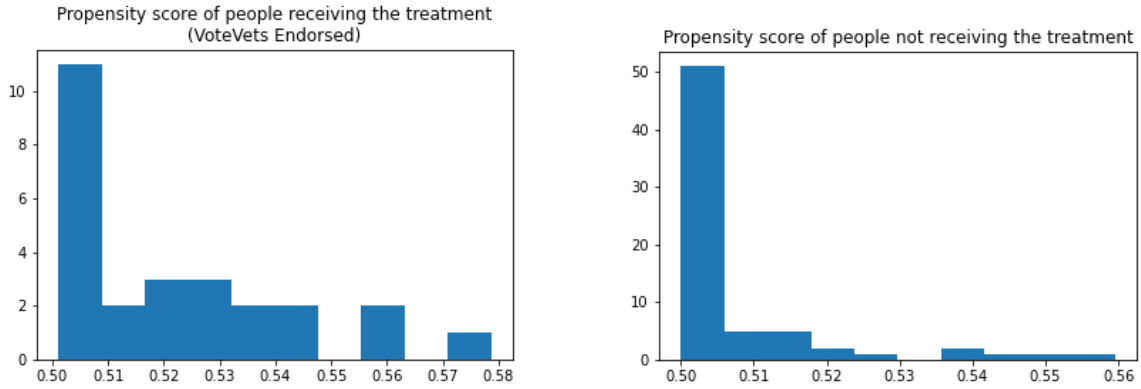
Figure 19. Histograms of propensity scores, grouped by treatment (Indivisible Endorsed)



Treatment 7: Receiving a VoteVets endorsement

Without accounting for confounders, the naive estimate is 0.408, which implies that a VoteVets endorsement increases the probability of winning by 40.8%. However, after applying inverse propensity weighting, the estimated average treatment effect and trimmed estimate both became 0.0744. After controlling for confounders, a VoteVets endorsement causes an 7.44% increase in the probability of winning the primary election with a 95% confidence interval of [-0.038, 0.194], which is not statistically significant.

Figure 20. Histograms of propensity scores, grouped by treatment (VoteVets Endorsed)



In order of highest to lowest causal effect, the treatments are: party support, an Emily's List endorsement, an Indivisible endorsement, a Justice Dems endorsement, being a Gun Sense candidate, a VoteVets endorsement, and an Our Revolution endorsement. Only the first three are statistically significant at the 95% confidence level. For the latter four that are not statistically significant, there is too much uncertainty to establish causality.

Unsurprisingly, establishment support and non-progressive endorsements had the greatest causal effect on election outcomes while endorsements from progressive organizations—namely, Justice Dems and Our Revolution—had less influence. Indivisible, which formed in response to Trump's presidency, likely had a statistically significant effect on primary outcomes because of the climate surrounding the 2018 election, when much of the focus was on the turmoil caused by Trump and the Republican majority in the House and Senate (Indivisible). Meanwhile, Emily's List aims to send pro-choice women to Congress, but they also invest a significant amount of money in their endorsed candidates (EMILY's List). While a portion of their success may be related to the tension regarding abortion rights in 2018, the dedicated support they provide to their endorsees likely paid off. The fact that the establishment takes precedence over progressive endorsements could indicate that people listen to establishment groups more, people resonate more with establishment policies, or that establishment groups have more to offer to their endorsees.

Discussion

Prediction with GLMs and Nonparametric Methods

Our GLM model had better interpretability of the results. For the purposes of explainability and understanding, our logistic regression model was better. However, the simplicity and interpretability of a logistic regression approach comes at the cost of accuracy: if the true functional form of the relationship is not in fact linear, even the best linear model will not be as accurate as other approaches.'In the context of accuracy, the decision tree performed best. We are not confident in applying our model to future datasets, because of the nature of open elections in America—only 19.6% of the eligible voter population voted in the 2018 primaries. If there was a different group of voters in future elections, this would impact the results. We would

only feel confident to apply our model to future data if voter demographics and public values stayed the same.

The performance metrics derived from the decision tree model imply that it fits better than the random forest, but could also be due to overfitting. This is indicated by the large number of splits and the depth of the trees, which was omitted as a result of the high dimensionality of our dataset. The random forest shows a good fit, albeit worse than the decision tree, with less risk of overfitting due to the ensemble nature of the model. Both models have recognized the importance of funding a candidate receives and whether or not that determines success from the election, which is seen in Figure 13.

We wanted to investigate the differences in scores for the decision tree and random forest. Typically an ensemble boosting method like random forests will perform better than a singular decision tree. Under both iterations of the model, we found the most significant feature was the 'receipts' column, which translates to the total funding a candidate received during their campaign, reported by the FEC. The feature significance score for the decision tree under ROS and RUS remained within the range (0.4, 0.6) consistently, which meant that this feature was always used, regardless of the sampling method applied. Naturally, a decision tree heavily influenced by one feature will tend to overfit the data, and because the random forest model prevents overfitting and selects a subset of features present in the data, it would make sense for the model to perform worse when one feature significantly impacts the outcome of a candidate winning the general election.

Both the decision tree and random forest aim to do classification of different candidates based on significant features. By limiting the depth of the trees classifying candidates, we are forcing the model to select splits that minimize the Gini impurity at each node, which is why it's important to finetune the model with hyperparameters such that it doesn't over classify the data. Regardless of the depth we set each tree, the selection of the most significant feature would always be the 'receipts' of a candidate, which is good for understanding the significance of this particular feature; however, a limitation of this model is the unavoidable effects of overfitting that we see as a result. Not only that, but our model consistently is reliant on the significance of one feature, which may not be the case for future elections.

Our dataset had an abundance of NaN or missing values that we dropped after one-hot-encoding our columns because they did not provide interpretability. Having data on these missing values would have improved our model and allowed us to interpret the results.

A limitation of Logistic regression models is that they are sensitive to multicollinearity where independent variables are highly correlated. This was an issue that we ran into when building our model because it leads to unstable coefficient estimates—we accounted for this using VIF.

In our model, four of the ten predictors' coefficients had a confidence interval that included 0. This means that in the presence of the other predictors, it is possible that there is no weight of that particular feature on the outcome variable. For example, whether or not the candidate was Emily Endorsed did not have a statistically significant effect on the log-odds of

the outcome. This introduces uncertainty in our model. While we noted a relationship between a candidate being endorsed by Emily's List in our EDA, the significance of this relationship is not reflected in our model. However, when we removed the four features that were not statistically significant we found that the AIC increased. We are uncertain whether these features affect the outcome or whether they are just noisy data.

Causal Inference

It is important to note that the variables used as confounders may not fully represent their intended meanings, which may skew the estimates. For instance, using 'Total Other Endorsements' as a proxy for popularity likely does not encompass the entirety of popularity's implications in this problem, which can lead to omitted variable bias. Overall, if our assumptions of unconfoundedness do not hold because our confounding variables are not comprehensive, then the IPW estimates will not exactly match the true estimates.

In addition to the three datasets used in this analysis, data that better encompasses popularity, such as social media following leading up to the election, would have been ideal. Moreover, a better metric for candidate partisan lean calculated in a similar way to the district partisan lean variable would have provided a better representation of how well each candidate matches their constituents' political views. If there had been less missing data for demographic features, the estimates would have been more accurate.

We are fairly confident that the causal relationship between party support and election outcome exists given the past research that aligns with this finding and the fact that we managed to include more extensive data on campaign finances and candidate views. Furthermore, a 22% increase in the likelihood of winning is quite large—even if we missed some minor confounders, there may still be a causal relationship. Based on the context of the 2018 US political scene—when Roe v. Wade was in jeopardy and many Democrats shared the goal of weakening Trump's influence—it makes sense that Emily's List and Indivisible also have a causal relationship on election outcomes. For the other treatments, we are rather confident that the causal effect is either too weak or nonexistent.

Conclusion

The research we have conducted examines the factors influencing Democratic primary wins in 2018. Key findings include the importance of gender, PAC endorsements, campaign funding, and STEM backgrounds in election outcomes, and these factors could impact a future election; however, our results, while insightful, may not be broadly replicable due to specific political conditions and voter demographics that may not be present in today's political climate. We were able to derive significant insights as a result of merging our three datasets, though this led to consequences of having a significant number of missing values and multicollinearity. In the future, we could explore changing political landscapes and voter behaviors over a time period instead of a static timeframe - nonetheless, we were able to observe the complexity of political data analysis and the influence of various factors on election outcomes.

Works Cited

Boudreau, C. (2016). The Persuasion Effects of Political Endorsements. https://www.democracy.uci.edu/newsevents/events/conference_files/boudreau_2016_politicalendorsements.pdf

Conroy, M., Nguyen, M., & Rakich, N. (2018, August 10). We Researched Hundreds Of Races. Here's Who Democrats Are Nominating. *FiveThirtyEight*. https://fivethirtyeight.com/features/democrats-primaries-candidates-demographics/

Desilver, Drew. (2018). Turnout in this year's U.S. House primaries rose sharply, especially on the Democratic side. https://www.pewresearch.org/short-reads/2018/10/03/turnout-in-this-years-u-s-house-primaries-rose-sharply-especially-on-the-democratic-side/

EMILYs List. *About Us*. https://emilyslist.org/about/

Federal Election Commission. (2018). *Browse Candidates.* https://www.fec.gov/data/candidates/?election_year=2018&office=H&party=DEM&is_active_candidate=true&has_raised_funds=true

FiveThirtyEight. (2022, December 19). *Primary Candidates 2018*. https://github.com/fivethirtyeight/data/tree/master/primary-candidates-2018

Indivisible. *About*. https://indivisible.org/about

Kamarck, E. & Podkul, A. R. (2018, October 23). Political Polarization and Congressional Candidates in the 2018 Primaries. https://www.brookings.edu/articles/political-polarization-and-congressional-candidates-in-the-2018-primaries/

Schaeffer, K. (2023, February 7). The Changing Face of Congress in 8 Charts. https://www.pewresearch.org/short-reads/2023/02/07/the-changing-face-of-congress/

Summary, B. (2010). The Endorsement Effect: An Examination of Statewide Political Endorsements in the 2008 Democratic Caucus and Primary Season. American Behavioral Scientist, 54(3), 284-297. https://doi-org.libproxy.berkeley.edu/10.1177/0002764210381709