Data 102 Final Project

Ziyi Ding, Laura Li, Nei Fang, Tina Chen

Introduction and Data Overview

In our project, we used the candidate information dataset (dem_candidate.csv and rep_candidate.csv) and the candidate spending data for both democratic and republican candidates (candidate_summary_2018.csv). Candidate information dataset is from FiveThirtyEight, a website that collects information about Democratic and Republican candidates in election results. The second dataset is from the Federal Election Committee, containing the disbursement summary of candidates. Both datasets are censuses, containing information about all candidates who appeared on the ballot in the Senate, House, and Governor primaries in 2018.

The candidate information dataset systematically excludes races of candidates expected for white candidates. Participants are fully aware that their information is collected and accessible online. Each row in both datasets represents a candidate, our results should be interpreted cautiously since our analysis cannot control for individual heterogeneity. At the same time, the candidate information dataset contains some measurement errors in race and veteran columns. According to the description of the dataset, the group creating the dataset manually checked each candidate's website to classify candidates' racial ethnicity and veteran status. This might result in data entry mistakes. We wish to have more information about candidates' gender, education level and geography information to incorporate into our model to better predict our first question of the election result. However, they are not available in our dataset.

Candidates summary data is from the Federal Election Committee, which represents a census. There is no information on how the data is obtained. However, because it is an official agency, we trust there to have no systematically excluded group, sampling bias, and measurement errors. Each row represents a candidate, their associated state and party, and disbursement, refund, contribution breakdown. We wish to have a uniquely identified column in candidate information datasets that makes references to candidate summary data's Cand_ID column.

Research Question

1. **Can we predict the election results of the Democratic candidates via their identities and endorsements?**

We would like to explore the above question as we are interested in what potential features are significant in predicting and affecting the final election results, which have real world implications on comprehending the types of candidates more likely to win or lose an

election. By answering this question, if our model exerts a good performance, we can predict the final election result to a certain accuracy based on a Democratic candidate's identity and endorsement information.

We decided to use the GLM and nonparametric approach since these models can be used for prediction. We will use the binomial GLM with a logit link function because our dependent variable is binary. We can interpret the coefficients and the confidence interval for the frequentist model and the credible interval for the posterior distribution of the bayesian model to determine what features are significant for this prediction task. For the nonparametric method, we choose to use random forest because all the features that we are using are categorical so the decision trees can split on these variables. Random forest is a good ensemble method that can reduce overfitting while achieving a considerably good accuracy.

2. **Does the amount of disbursement candidates spend affect their results in elections?**

We aim to establish causal relationships between the total disbursement amount and final election results for Democratic and Republican candidates separately, and juxtapose the difference between expenditure, party affiliation and final election result. By studying this question, we are able to understand the role of disbursement in elections. Investors and campaign programs may quantify the impact of their monetary investment with our research result. At the same time, answering the question can have further implications on candidates' decision makings as well as their campaign strategies.

We decide to use inverse propensity weighting (IPW) to account for confounding variables. IPW is an appropriate way to eliminate the effects of large numbers of confounding variables, and evaluate the causal relationship between disbursements and final election results.

EDA and Data Cleaning

For the first question, we are only working with the democratic candidates dataset since the republican dataset does not have sufficient attributes relating to the candidate's identity. Since we are investigating the election results, we used both the Primary Status and Primary Runoff Status to construct the Final Election Result column based on the following conditions:

- if Primary Status == 'Lost' then 'Lost'
- if Primary Status == 'Advanced' and Primary Runoff Status == 'None' then 'Won'
- if Primary Status == 'Advanced' and Primary Runoff Status == 'Advanced' then 'Won'
- if Primary Status == 'Advanced' and Primary Runoff Status == 'Lost' then 'Lost'

We removed the rows where Primary Status is 'null' or the Primary Runoff Status is 'On the Ballot' since we won't be able to know the results of the election. For the endorsement columns, we looked at the number of NaN values for each column and removed the column "No Labels Support?" since it has too many missing values. Then we dropped some irrelevant columns and selected only the identity-related and endorsement columns. For the selected endorsement columns, we filled the NaN values with No, since according to the data description, the

endorsement field is ‘Yes’ if the candidate is endorsed, ‘No’ if the candidate is running against an endorsed candidate, and ‘NaN’ otherwise. Hence having an empty value is equivalent to not being endorsed by that party.

In the first question, we engineered the column final_election results for democratic dataset and republican datasets and stored them as dem_original_with_results and rep_original_with_results. For the second question, because we need to merge candidate_summary_2018.csv (stored as disbursement) with those two datasets based on candidate name information, we first lowercase everything in cand_name, then split up each name based on commas. After examining the data and comparing with politicians’ names online, we identified that the first word is Candidates’ last name and the second word is Candidates’ first name. Therefore, we reordered the Cand_name so it contains the first name, then followed by a whitespace and then the last name. Now disbursement.Cand_name displays the same structure as dem_candidates.Candidates and rep_candidates.Candidates.

Then, we found the rows in disbursement where “Cand_Party_Affiliation” is “DEM”, and left-merged dem_original_with_results with it using a rapid fuzz package using an accuracy cutoff of 86 out of 100, preserving the single match with the highest score. Afterwards, we did another round of matching for candidates left out by the rapid fuzz. In particular, we looped through these candidates and checked if both their first and last names are uniquely contained by an entry in disbursement.Cand_name, prior to the name update mentioned above. If so, we updated the information accordingly.

After these two steps, we ended up with 668 matched rows for the democrats and 630 matched rows for the republicans. We then selected these columns from both merged datasets: 'Candidate', ‘Total_Receipt’, 'Total_Disbursement', 'Exempt_Legal_Accounting_Disbursement', 'Fundraising_Disbursement', 'Other_Disbursements', 'State', 'Final Election Result', and 'political_affiliation', and concatenated the two datasets, referred to as match. Finally, when producing the visualizations, we found the interquartile range (IQR) for Total_disbursement column in match and considered only rows with Total_disbursement column below 1.5 IQR above Q3. Because there were no values below 1.5 IQR below Q1, the above predicate is sufficient to reduce outliers.

Visualizations

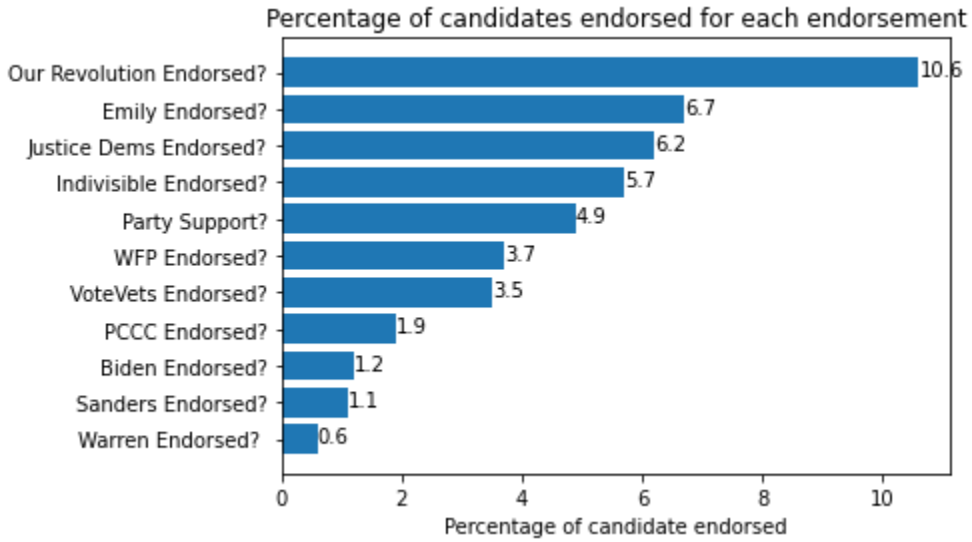

Figure 1. Percentage of Candidates Endorse for Each Endorsement

This bar plot shows the percentage of endorsed candidates for each endorsement in Democratic parties in a descending order. Based on the visualization above, we can observe that One Revolution Endorsed the highest proportion of candidates, a bit over 10%, whereas Senator Elizabeth Warren endorsed the least percentage, less than 1%. Hence, we can infer that different people or organizations may have different standards when it comes to which candidate they endorse. For instance, we can observe that individuals, including Biden, Sanders, and Warren, are less likely to endorse candidates than organizations such as Our Revolution. In order to answer the question about the relationship between endorsement types and final results, we can take these endorsement percentages into account when creating our model.

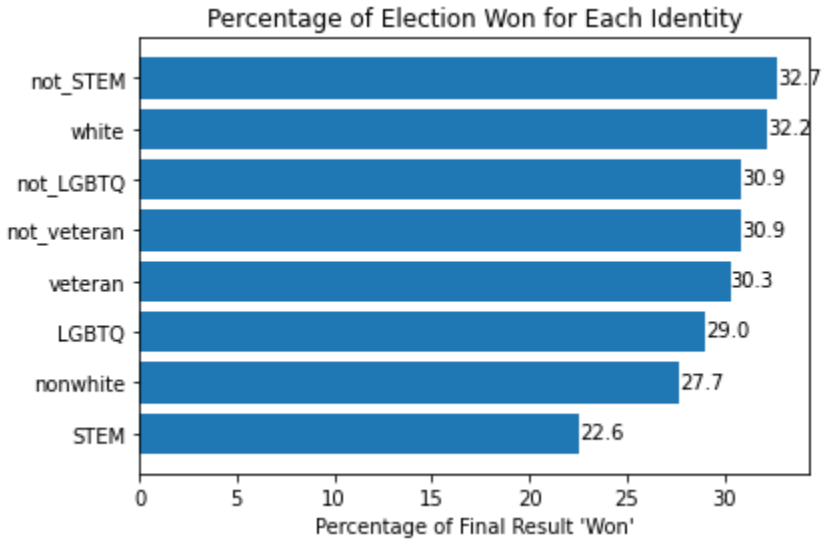

Figure 2. Percentage of Election Won for Each Identity

The visualization above shows the percentage of winning results for each type of "identity" in the dataset in descending order. The listed identities are the only identity types available in the dataset. The categories "not_STEM" and "white" have the highest winning rate, with around 32 percent of the democrat candidates in the dataset obtaining a winning result. On the other hand, the "nonwhite" and "STEM" categories have the lowest winning rate, with approximately 27 and 22 percent of the democrat candidates in the dataset obtaining a winning result respectively. Thus the identities have some observable differences according to the visualization above. It can potentially be a useful variable for predicting the final election result for a given candidate. Whether these differences are significant will be further evaluated in the model section.

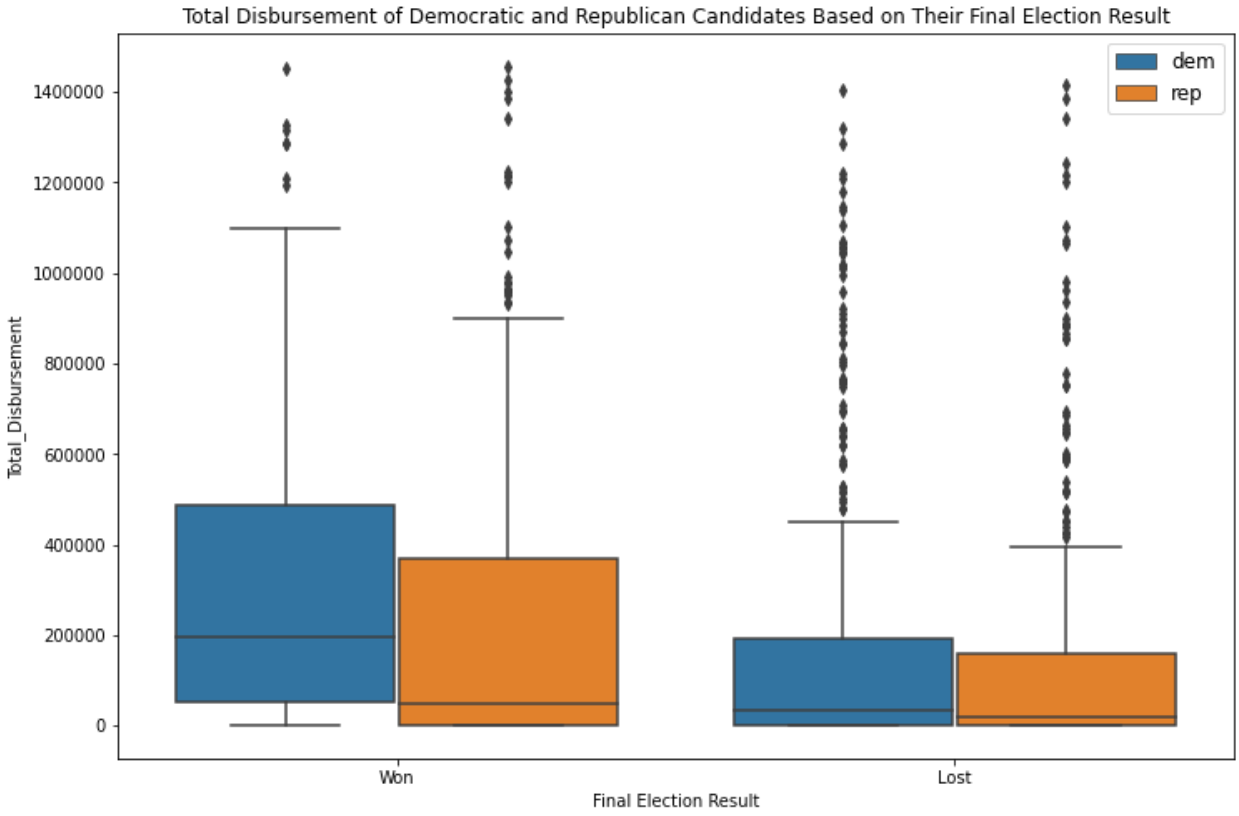

Figure 3. Total Disbursement of Democratic and Republican Candidates Based on Their Final Election Result

Figure 3 indicates that total disbursement of candidates who won the election are slightly higher than those who lost. For Democratic and Republican candidates who lost in the final election, their median disbursements are below 100,000 dollars. The difference in median total disbursement between democratic and republican candidates who won the election, which will be further investigated in the causal inference section. In addition, the variation of total

disbursement among the winning group is much greater than the variation of total disbursement among the losing group. A noteworthy aspect of the visualization is the outliers among both groups. Although we already cleaned the outliers based on interquartile range, there are still many high total_disbursement candidates. This indicates there are some candidates that are willing to invest significantly more funding.

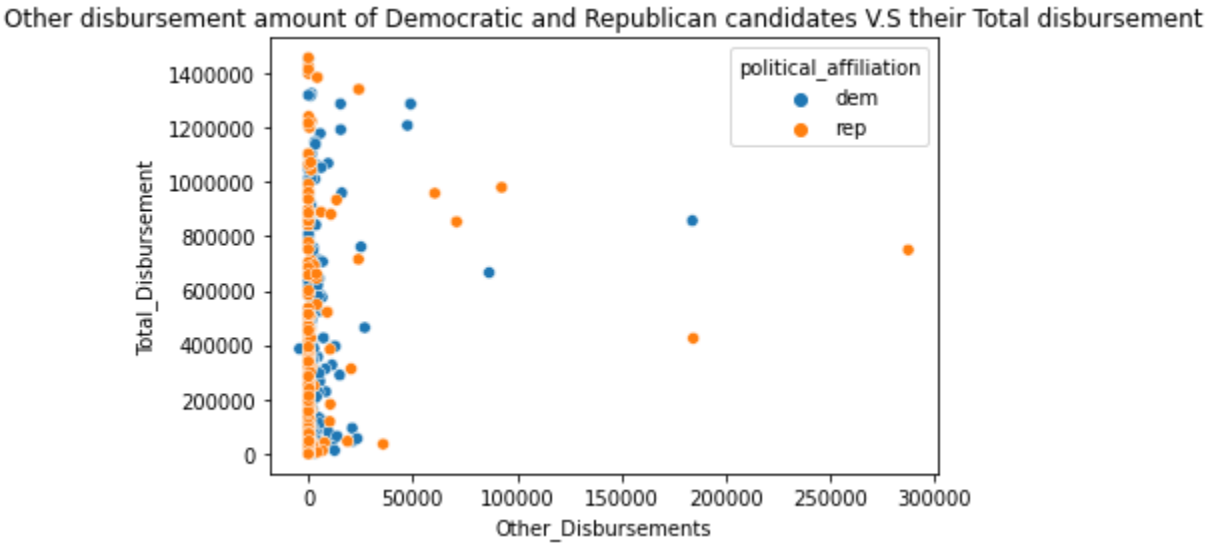

Figure 4. Other Disbursement Amount of Democratic and Republican Candidates V.S Their Total Disbursement

The scatterplot above demonstrates the relationship between other disbursements and total disbursement. As we can observe from the trend above, the other disbursement amounts are clustered around 0. Hence we can infer that the other disbursements variable is not as important as compared to total disbursements. We will take this into account when developing the causal inference model.

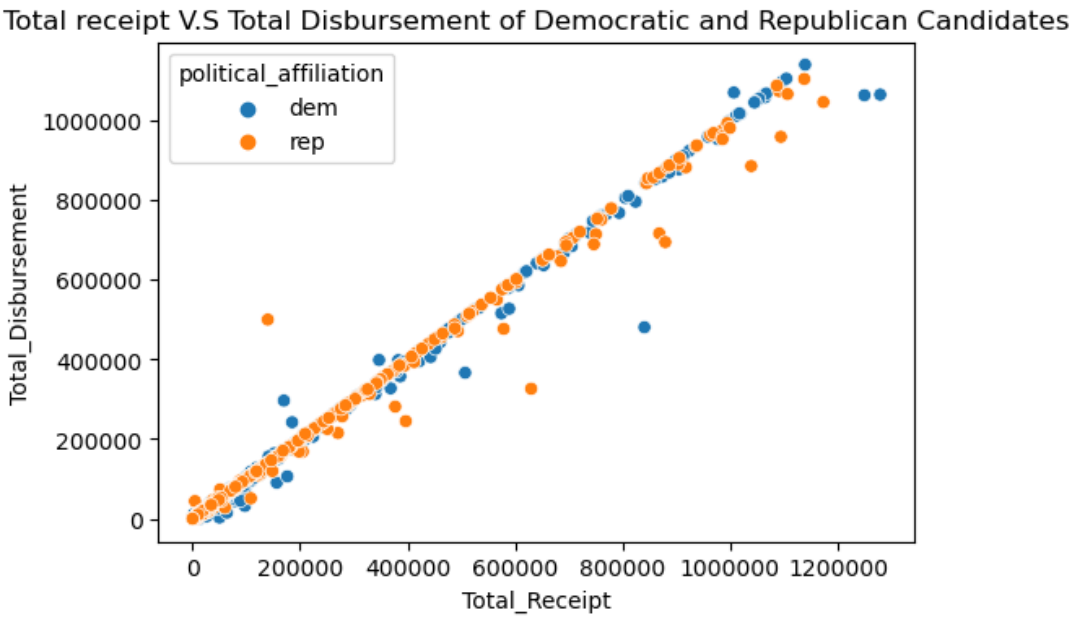

Figure 5. Total Receipt V.S Total Disbursement of Democratic and Republican Candidates

This scatterplot above demonstrates the relationship between total receipt and total disbursement. We can observe that there is a mostly linear relationship, with only a few outliers. According to the diagram, spending less than received occurs more frequently than over-disbursements. Due to this high consistency between disbursements and receipts, we could use disbursements as a tool to explore the causal relationship between financial investments and the result of elections.

Prediction with GLMs and Nonparametric Methods

Methods

Feature selection

To predict the final election result for a candidate given their identity and endorsement, we use the features including 'Race', 'Veteran', 'LGBTQ', 'STEM', 'PartySupport', 'EmilyEndorsed', 'BidenEndorsed', 'WarrenEndorsed', 'SandersEndorsed', 'OurRevolutionEndorsed', 'Justice Dems Endorse', 'PCCCEndorsed', 'IndivisibleEndorsed', 'WFPEndorsed', and 'VoteVetsEndorsed' to predict the 'FinalElectionResult'. These features represent the candidate's identity and whether they received certain endorsements which relate directly to our research question. Moreover, these features are binary which can all properly fit into our GLM and random forest model.

GLM justification and assumptions

We will be using the binomial GLM with a logit link function which ensures that the prediction is positive between 0 and 1 as the election result cannot be negative. We choose to use logistic regression since the Final Election Result variable is discrete and binary. Since we have no prior knowledge about the relationship between identity and endorsement and the election result, we used the default flat and uniform prior for our Bayesian GLM. For the models, we assume that there is a linear relationship between the dependent variable and the independent variables after applying the inverse link function. We also assume homoscedasticity and little or almost no multicollinearity between the explanatory variables. Lastly, we assume the explanatory variables are independent of each other, where one type of identity or endorsement will not influence another.

Nonparametric justification and assumptions

We choose to use random forest for the nonparametric method because all the features that we are using are categorical so the decision trees can split on these variables. Random forest is good for classification tasks and we choose it over a simple decision tree because we have many predictive variables and we want to reduce overfitting and variance. Using the random forest model, we assume that each variable has only one level and there are no missing values in the input data since we removed it during the data cleaning phase.

Model evaluation methods

For the Frequentist GLM model, we will evaluate the 95% confidence interval and p-value for each coefficient and for the Bayesian GLM model, we will evaluate the 95% credible interval (HDI) for each coefficient. If a feature's interval contains 0, then that indicates that the feature is not predictive for the election result. If the p-value is smaller than 0.05, then we consider the feature as significant for predicting the final election result.

To evaluate each model's performance, we used a train test split with 80% training data and 20% testing data. Both RMSE and accuracy will be calculated. We will compare each model's RMSE as well as accuracy score to evaluate how well each model fits the data and generalizes it to future data. Both metrics are calculated based on a 0.5 cutoff for GLM and random forest.

Frequentist GLM Model Implementation

Results

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Generalized Linear Model Regression Results

Dep. Variable:	FinalElectionResult	No. Observations:	642
Model:	GLM	Df Residuals:	626
Model Family:	Binomial	Df Model:	15
Link Function:	logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-334.49
Date:	Mon, 05 Dec 2022	Deviance:	668.97
Time:	22:33:59	Pearson chi2:	638.
No. Iterations:	22		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-1.3798	0.171	-8.052	0.000	-1.716	-1.044
Race	0.4006	0.197	2.029	0.042	0.014	0.788
Veteran	-0.0278	0.311	-0.089	0.929	-0.638	0.582
LGBTQ	-0.1299	0.469	-0.277	0.782	-1.048	0.788
STEM	-0.4770	0.273	-1.745	0.081	-1.013	0.059
PartySupport	2.9232	0.784	3.727	0.000	1.386	4.460
EmilyEndorsed	1.7401	0.456	3.816	0.000	0.846	2.634
BidenEndorsed	22.3076	2.14e+04	0.001	0.999	-4.19e+04	4.19e+04
WarrenEndorsed	22.8350	3.21e+04	0.001	0.999	-6.29e+04	6.29e+04
SandersEndorsed	1.2921	0.851	1.518	0.129	-0.377	2.961
OurRevolutionEndorsed	-0.0680	0.339	-0.200	0.841	-0.733	0.597
JusticeDemsEndorsed	0.1326	0.421	0.315	0.753	-0.693	0.958
PCCCEndorsed	0.3075	0.831	0.370	0.711	-1.321	1.936
IndivisibleEndorsed	1.6979	0.391	4.348	0.000	0.932	2.463
WFPEndorsed	0.4983	0.473	1.053	0.292	-0.429	1.426
VoteVetsEndorsed	0.2272	0.606	0.375	0.708	-0.961	1.416

Figure 6. Frequentist GLM Training Result

Based on the result above, Race, PartySupport, EmilyEndorsed, and IndivisibleEndorsed are significant features that are predictive of the final election result since the p-value of their coefficients are small (below the 0.05 threshold). For the other variables, their confidence intervals all include 0 and their p-value does not indicate any significance. We can thus conclude that for features other than Race, PartySupport, EmilyEndorsed, and IndivisibleEndorsed, they are insignificant in predicting the candidate's final election result.

To interpret the uncertainty of the frequentist GLM model, we again refer to the 95% confidence interval listed above for each coefficient. For example, for the Race variable, we are 95% confident that a white candidate will increase the log odds ratio of winning between 0.04 and 0.737. For the PartySupport, we are 95% confident that if a candidate receives support or endorsement from the party, then the log odds of winning would increase by between 1.834 and 4.83. The RMSE and accuracy score will be further interpreted in the discussion section.

Bayesian GLM Model Implementation

Results

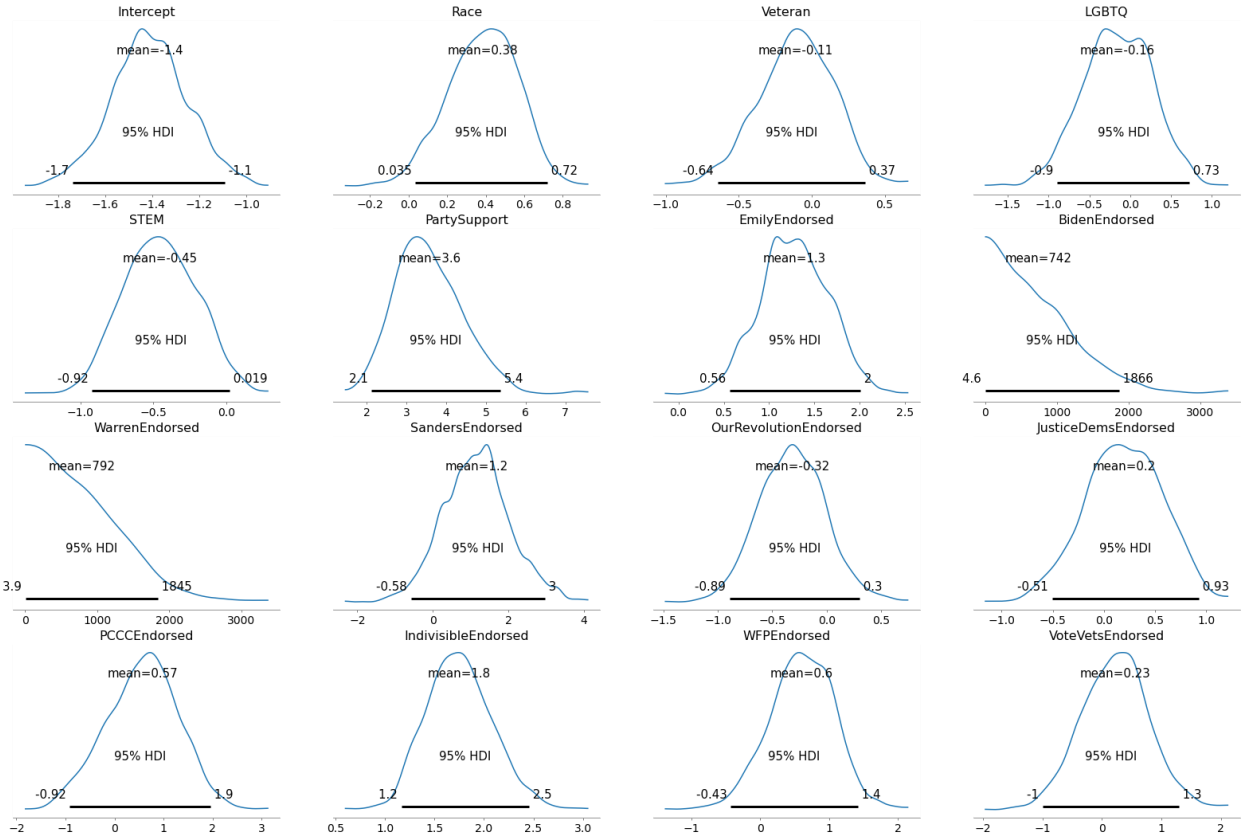

Figure 7. Bayesian Modeling Coefficient Posterior Distribution

In the Bayesian model, since we don't have any prior information about the data, we use the uniform prior and generate 100 samples to draw their posteriors. In the figure above, the x-axis indicates the coefficients of each categorical variable and the y-axis represents the corresponding probability density of the posterior distribution. We then set a 95% highest density interval to each subplot. This means that there is a 95% probability that the coefficient of the posterior feature would lie within the interval, given the evidence provided by our data. If the HDIs include 0, this indicates that the corresponding features doesn't associated with the final election result. Based on our observations, Race, PartySupport, EmilyEndorsed, IndivisibleEndorsed are all positively correlated with the final election result. Moreover, BidenEndorsed and WarrenEndorsed are skewed to the right, showing a large variance. The RMSE and accuracy score will be further interpreted in the discussion section.

Nonparametric model - Random Forest

Results

- Training RMSE: 0.4517187667528992
- Testing RMSE: 0.479388844880931
- Training Accuracy: 0.7959501557632399
- Testing Accuracy: 0.7639751552795031

We split the dataset into a training set and testing set and compute the root mean squared error and accuracy for each of them in a random forest model. Since the difference in RMSE and accuracy between the training set and test set are small, this shows the training model predicts the testing set well.

Discussion

Model Performance

	Frequentist GLM Model	**Bayesian GLM model**	**Random Forest**
RMSE (root mean square error)	**Training:** 0.48175	**Training:** 0.41396	**Training:** 0.45172
	Testing: 0.46625	**Testing:** 0.47287	**Testing:** 0.47939
Accuracy	**Training:** 0.76791	**Training:** 0.77086	**Training:** 0.79595
	Testing: 0.78261	**Testing:** 0.77640	**Testing:** 0.76398

We calculated the training and testing RMSE and accuracy to evaluate each model. We concluded that the Frequentist GLM model performs better than the Bayesian model and Random Forest model because it has the lowest testing RMSE and the highest testing accuracy. However, their performance does not differ significantly since all metrics for each model are relatively close to each other. We are confident in applying our best model to future datasets because the difference between the training and testing RMSE and accuracy is small. Notably, the testing RMSE is even smaller than the training RMSE and the testing accuracy is greater than the training accuracy, indicating that the model has good generalizability. Although the deviance of our model is large, the log-likelihood is reasonable. To account for this, we can further improve our model with more predictive features which will be discussed in the following sections.

Model Fit

To evaluate how well each model fits the data, we first looked at their training accuracy. The frequentist, bayesian and random forest models have an accuracy around 76 percent, 77 percent, and 79 percent respectively. Comparatively, the random forest model best fits the training data based on accuracy. However, we concluded that our models do not fit each model very well since the RMSE are around 0.45 which is large relative to the range of outcome variables which is between 0 and 1.

Difference Between Bayesian and Frequentist GLM Implementations

One difference between the Frequentist and Bayesian model is that the PartySupport variable has relatively large standard error but the posterior distribution of PartySupport in the Bayesian model does not display a wide range of values. On the other hand, the posterior distribution of BidenEndorsed and WarrenEndorsed have a large spread around 3000 which might be caused by significant outliers. This is also roughly captured by the Frequentist model, where both the coefficient and the standard error of BidenEndorsed and WarrenEndorsed are relatively much larger than those of the other variables.

Interpretation of Results

In our result section, we have interpreted the results of all three models. As we discussed above, Race, PartySupport, EmilyEndorsed, and IndivisibleEndorsed, are significant in predicting the candidate's final election result for the frequentist model. Based on the Bayesian model, these are also the features that have correlations with the final election result. Moreover, these features are all positively correlated with the dependent variable based on both models and the other features are insignificant. We used confidence intervals and credible intervals to quantify our uncertainties. Both intervals show similar ranges for each coefficient, which validate our thoughts about the influence of some features on the election results.

Limitations

The limitation of the GLM model is that we assume uniform distribution due to lack of knowledge about the prior distributions. Both GLM models assume homoscedasticity and little or almost no multicollinearity between the explanatory variables, but in reality, this is nearly impossible. There are connections between parties, and some endorsed parties will sponsor candidates in groups.

For non-parametric models, it is hard to interpret the outputs. Random forest is an ensemble method, the inherent meaning might get lost in the decision making process, which makes it hard to present the result to the audience. Since the model is formed by an arbitrary

combination of features within the “black-box” model, it is hard to foster further meaningful analysis.

Improvements

To improve our models, we should incorporate more identities for each candidate. For example, under race variable, there are only three categories: white, non-white, and null. In reality, there are more racial groups, so if we can join another dataset about the race of the candidates, we will better analyze how race impacts the final election results. Other identities, such as religion, geographic locations, and education levels can also be crucial factors that will influence the election results. Due to time limitation, we only used zero to replace the null values in endorsement types. If we have a dataset with more detailed information about types of endorsements, we will have lower variances when predicting the endorsement coefficients.

Causal Inference

Methods

Treatment and Outcome Variables

Since we are trying to establish the causal relationship between disbursement and results in elections. The units are individual candidates, and the treatment is having a disbursement amount more than a threshold. The threshold is calculated corresponding to the percentage of candidates in each party who won the election. For the democrats, the threshold is 386569.2 and for the republicans, the threshold is 147713.8. We consider the candidates whose disbursement amounts are higher than their respective threshold values to be treated. The outcome is the results of the election, a binary variable.

Confounders and Adjustments

We are aware that candidate information, such as their identity and endorsement received, are confounding variables. There are some common confounders among candidates from the two parties, including the state they are running in and the position they are running for. They affect both the treatment and the outcome as geographic and position factors will affect the amount of money spent, and also affect the results of elections since each position is different. Both datasets contain information on endorsements but coming from different sources. In addition, for the Democratic datasets, confounders include identity information including ‘Race’, 'Veteran?', 'LGBTQ?', 'Elected Official?', 'Self-Funder?', 'STEM?', 'Obama Alum?', and 'Party Support?'.

This information affects both treatment and outcome: for example, candidates that received multiple endorsements are more likely to receive more funding, from high public exposure and popularity, which they can spend. On the other hand, they are also more likely to

win elections due to the same reason. Similarly, attributes that detail candidate identity can also be considered as confounders. For instance, candidates with military backgrounds can better appeal to the veteran communities, and subsequently receive more funding from them. They are also more likely to receive votes from those communities. All binary confounding variables use 1 for 'Yes' and 0 for 'No'. Non-binary confounding attributes such as State are one hot encoded. After these feature engineering steps, we transformed all confounding columns into a Numpy array and defined it as X, a matrix of confounding information.

In order to adjust for the confounders, we used inverse propensity weighting. We first calculated the $\widehat{e}(x) = P(Z = 1|X = x)$ using logistic regression. Then, we removed the outliers and only used the candidate data with a propensity score of $0.1 \leq \widehat{e}(x) \leq 0.9$. Lastly, we estimated treatment effect as $\tau_{IPW} = \frac{1}{n}\sum_{i=1}^{n}(\frac{z_i y_i}{\widehat{e}(x_i)} - \frac{(1-z_i)y_i}{1-\widehat{e}(x_i)})$, where the x_i are features, y_i are the outcomes, and z_i are the treatments. For reference, we also calculated the naïve estimate which does not take the confounders into account: $\tau_{naïve} = \frac{1}{n_1}\sum_{i=1}^{n_1}(z_i y_i) - \frac{1}{n_0}\sum_{i=1}^{n_0}(1 - z_i)y_i$, where n_1 is the number of treated candidates, and n_0 is the number of untreated candidates.

Results

For Republicans, our naive estimate result is approximately *0.231*, and the inverse propensity weighting estimate of treatment effect is approximately *0.128*. Hence, total disbursement does cause higher likelihood of winning the election among Republican candidates. After accounting for confounding variables, such positive casual relationships still exist, but the power of the relationship is slightly weaker. A different trend is observed for democrats. The naive estimate is slightly lower than our inverse propensity weighting estimate, approximately *0.432* and *0.457* respectively. The statistics are close to each other, demonstrating that a moderate casual relationship between total disbursements and final election results can be established among democratic candidates. Juxtaposing the treatment effects of both groups, we see that disbursement amounts have a much stronger impact on democratic candidates than republican candidates. This result matches Figure 3, where Democratic candidates have greater difference in median total disbursement between winning and losing groups than Republican candidates. From the visualization, Democratic candidates are more willing to invest more money for their candidacy. This pattern is rational because their total disbursements have a greater impact on their final election results.

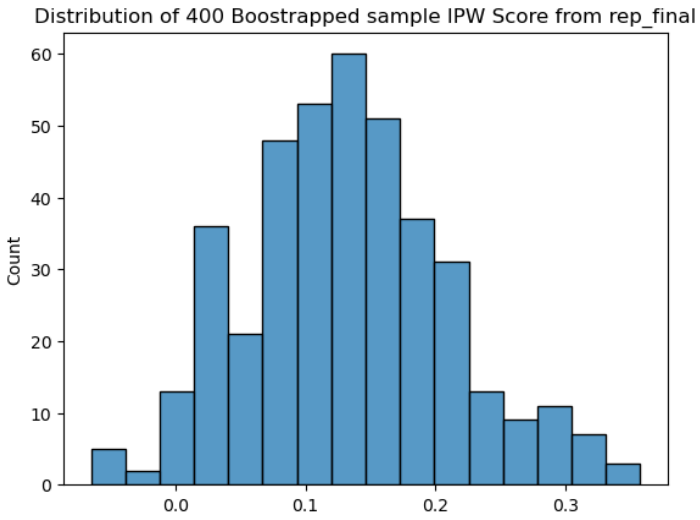

Figure 8. Distribution of 400 Bootstrapped Sample IPW Score from rep_final

The 95 percent confidence interval of the treatment effect generated by bootstrapping the republican data is [-0.024, 0.289]. The calculated treatment effect of 0.128 is included in the interval. Since 0 is also in this interval, this further supports the claim that disbursements don't have a strong impact on Republican election results.

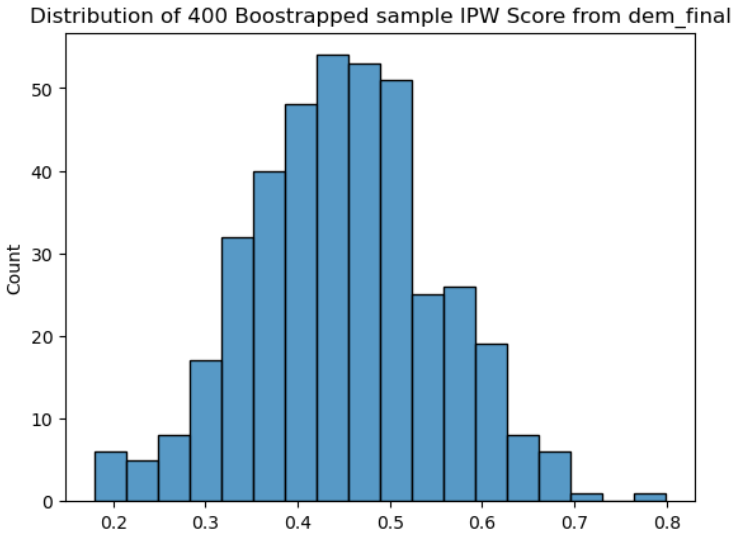

Figure 9.Distribution of 400 Bootstrapped Sample IPW Score from dem_final

The 95 percent confidence interval of the treatment effect generated by bootstrapping the democratic data is [0.247, 0.652], which includes our estimated treatment effect of 0.457. As compared to our Republican samples, we see that treatment effect does not include 0 and have higher mean statistics, meaning that total disbursement has a stronger influence on Democratic election results.

Discussion

The first limitation is that we attempted to string match the Democratic and Republican candidates in their respective datasets with disbursements dataset. After string matching with rapidfuzz, we manually assigned a few rows with missing candidate names based on domain knowledge. This process may lead to candidate mismatch. Unmatched candidates are also discarded, thus losing some information.

Secondly, inverse propensity weighting assumes that there are no unmentioned confounders. Nevertheless, it is possible that other confounding variables, unlisted in our dataset, exist. These confounders would undermine the legitimacy of our propensity score, hence becoming a limitation of our method.

As discussed previously, there are some other confounders that are not taken into account in our IPW. Hence, it would be great if we can get additional, more detailed data on the candidates themselves. For instance, the "Race" column is unclear with Null values in the democratic dataset, and it is not included in the republican one. Other information such as gender, level of education, and previous government experience are all confounding variables in this case. Having additional data addressing this information can help us improve our IPW approach and help us to better establish the causal relationship. In the end, as previously discussed, we are confident that there is a causal relationship between total disbursements and final election results, especially for the democratic candidates.

Conclusion

Summary of Key Findings

For our first research question, we used prediction with GLMs and nonparametric methods and concluded that Race, PartySupport, EmilyEndorsed, and IndivisibleEndorsed are significant features that can be used in predicting a Democratic candidate's final election result. And we used causal inference to answer our second research question. We found that there is a causal relationship between total disbursements and final election results, and the causal effect is stronger for the democratic candidates.

Generalizability of Results

Based on our model performance on test data, we conclude that our model can be generalized to predict on future datasets with similar information. However, since our data only apply to Democratic candidates who ran in the House and Senate primaries in 2018, they are specific and thus difficult for generalizing to other parties and years. Similarly, our causal inference is established based on data exclusively in 2018. Although we believe elections follow

similar trends year by year, our treatment effect level would be less accurate for Republican and Democratic candidates in another year as well.

Call to Action

In light of our findings, Democratic candidates can seek Party Support, Emily endorsement, and Indivisible endorsement to potentially increase their chance of winning. The other types of endorsement that were deemed insignificant by our model can be treated with lesser importance. Based on causal inference results, Democratic candidates and their affiliated campaign should prioritize obtaining more funds and increase their disbursement more than Republican candidates as a campaign strategy since their final election result is more affected by disbursement amounts.

Merging

For the first research question, the dem_candidates.csv has all the information we needed for answering our question so we did not merge another data source. For causal inference, we merged candidate_summary_2018.csv with dem_candidates.csv and rep_candidate.csv based on candidate name using string matching, which enabled us to link the election result of each candidate to their disbursements. One downside of merging is that we lost some candidate information since the names of the candidates cannot be matched perfectly.

Limitations

In our first question, our dataset is only based on the election result in 2018. If we have a wider range of data or the election outcomes in different years, our model will be more generalizable. In our second question, rows that have low string match accuracy and cannot be matched based on our domain knowledge are discarded. Therefore, our analysis cannot account for candidates' information in omitted rows.

Future Studies

For our first research question, we only incorporate the data in 2018. For our next step of research on this topic, we could possibly check out future years' election results and validate our conclusion of whether the features that we select also correlate with the election result in other years. Moreover, we can also look at more specific types of identity to make our features more accurate in predicting the election outcomes. For our second research question, future research may reach a more generalizable conclusion if we have access to every candidates' disbursement information in different years. In addition, future studies can also infer causal relationships of other variables to final election results.