Data102 Final Project

Michael Chien, Sarah Shikanov, Wenli Fei, James Marquez

December 2021

1 Introduction

1.1 Data Overview

The democrat and republican data were generated by sampling. This was taken from the class dataset. At quick glance from the our pre processing step below we can see that in our first entry we have a governer running from office for Alabama. At quick glance we can see that the Partisan Lean is heavily negative which means its heavily red. This is consistent with what we would expect from before we began doing anything. We are not worried much about bias as this data is simply a representation of endorsements and the primary. Important features that would be useful if there was a way to get the total contribution with the candidate. We tried merging the dataframes with another dataset but it was very non trivial and thus decided to work with what we had. This dataset is public so all participants are aware of this dataset. As to the granularity each row represents a candidate running for office with endorsements and primary election data.

The 2016 Polling dataset was also generated by sampling. This was a dataset from Kaggle and we decided to use this because the election dataset given only covered endorsements and there weren't very many cool problems we could think of analyzing. Again by looking at the data we can get a quick glance at the polling advantages for each candidate and they seem to be consistent.The participants are aware of the collection and use of the data as it's published by pollsters. We can see unlike the class dataset the polling dataset is worrisome with selection bias due to the unreliability of some pollsters but there is a column for grade which should help us identify good pollsters. The columns are sufficient in this dataset to perform what we want in building our model.

1.2 Research Questions

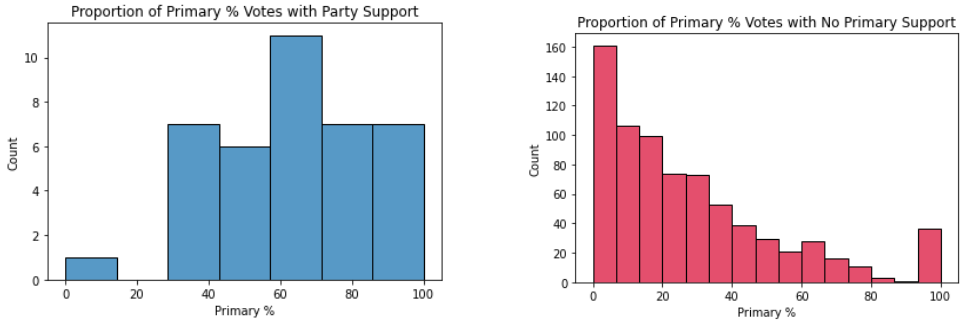
1. Our first research question is *can we predict who will win the primary elections based on number of endorsements plus any good features we can find?* By answering this question we can answer whose endorsements have the strongest influence and what are the defining features that can help a candidate win an election? Non-parametric methods/GLMS are perfect for this question because we can feature engineer features to help us predict who will win the election. We can then measure the accuracy of our model and reflect on our choices for our model.

   We use GLMs because our data is binary we anticipate that we can find a link function in our GLM which will allow to exploit the linear structure of our data to use logistic regression in our predictions.

2. Our second research question is *can we estimate the poll leads between Clinton and Trump during the 2016 presidential election?* This question is easy to answer because by estimating poll leads we can gauge overall sentiment throughout our country using polls which can then later be analyzed for future election data. We use Bayesian Hierarchical modeling as using Bayesian Inference can helps us estimate our posterior distribution and we will then create a graphical model using the regions of the country.

1.3 EDA

In our EDA data cleaning primarily involveed filling our null values with 0 as we considered this not an endorsement. For our primary election dataset we first explored the possibility of good features. We looked at party support as we believed most candidates would rely on party support



(a) (b)

Figure 1: Party Support Histogram

From our histograms we can see that party support is quite important. The histogram with no party support is right skewed with many of the votes not even above 50 percent. Out of 750 candidates there is a few percentage that is above 50 percent compared to the first histogram. We should keep this in mind when building our model. We then took a look at a map of the US with partisan leans.

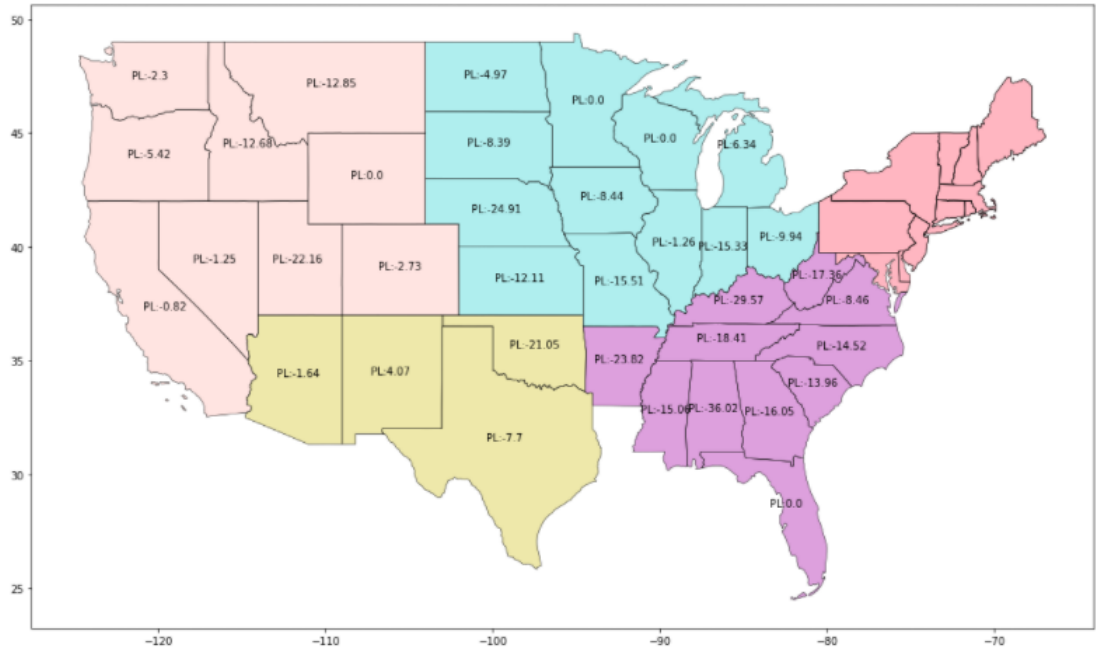
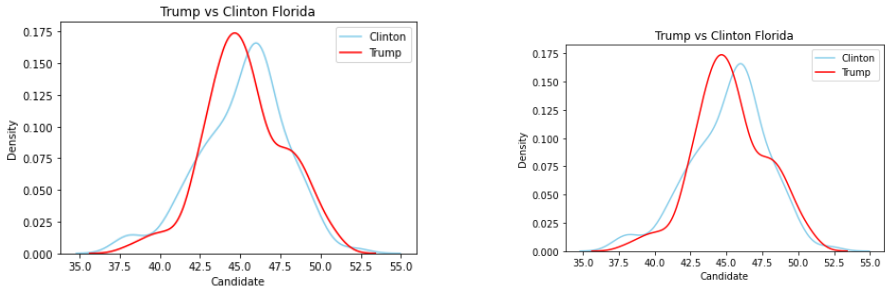


Figure 2: US Maps Partisan Lean

This map is a useful visualization of the partisan leans between states. We see that the South is heavily republican while the West is less republican and leaning towards more Democrat. This is quite as expected but can be useful when performing Bayesian Inference. For the sake of brevity we omitted an image of the northeast region as it required another image but it is consistent that is is more Democratic in partisan lean.

These visualization motivate our research question because it allows to get a better idea of a better feature to use in this case party support and it allows us keep our beliefs of the region differences consistent.

Next we look at our polling data from Kaggle. We first want to analyze the states as this will allow us to see where the leads the most contentious as we want to build our hierarchical model later using state regions.



(a) (b)

Figure 3: Trump vs Clinton Battleground

We can see for battleground states even though a state has a higher mode such as Clinton in Florida, Trump in the polls still has the lead as his avergae poll rating is higher. This is most likely due to that Trump has a higher density in Florida giving him a slight lead. Most of the differences come from the distributions being slightly wider or taller in key areas. We then introduce a new variable, *adjusted lead*, which will be the difference between Clinton and Trump poll ratings where a positive value means Clinton has a lead over trump and negative means Trump has the lead over Clinton.

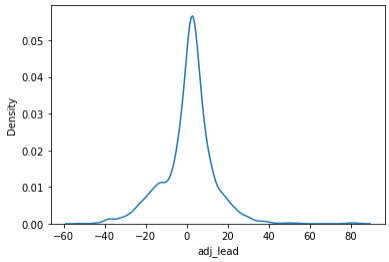


Figure 4: Adjusted Leads Distribution

These visualizations are helpful for the polling dataset because in our Bayesian Hierarchical model these will be useful in constructing our priors as the above visualization helps us consider a uniform distribution.

2 Non-parametric Methods and GLMS

2.1 Methods

For our non-parametric methods we will attempt to predict who won the primary based on the number of endorsements and the proportion of voters. We will need to feature engineer our data to count the number of endorsements along with the proportion of people. We first consider logistic regression. Then we will see if there's a better model for us to use.

Again we are trying to predict the same thing as above in our nonparameric methods. We will be using a binomial glm because we have count data and our link function should work for our binary data. We will then evaluate the model by first using a frequentist model and analyzing the deviance and Pearson chi-squared before than analyzing the posterior distribution.

2.2 Results

For our non-parametric methods we first had a 70 percent accuracy using logistic regression. We concluded that our data must not be very linearly separable and thus we will use a random forest. Our random forest accuracy was much better at 81 percent. However, our training accuracy was much higher in our random forest then our test set accuracy which means we must've over fitted our data.

```
train_accuracy = accuracy_score(train['pred'], train['Won Primary'])
test_accuracy = accuracy_score(test['pred'], test['Won Primary'])
train_accuracy, test_accuracy
```

```
]: (0.7376014427412083, 0.7016806722689075)
```

(a)

```
from sklearn.metrics import accuracy_score

train_accuracy = accuracy_score(train['tree_pred'], train['Won Primary'])
test_accuracy = accuracy_score(test['tree_pred'], test['Won Primary'])
train_accuracy, test_accuracy
```

```
(0.8151487826871056, 0.6932773109243697)
```

(b)

Figure 5: Accuracy for our nonparametric methods

Lets see if we can do better with our GLM. We first consider a frequentist GLM.

Generalized Linear Model Regression Results

Dep. Variable:	Won Primary	No. Observations:	1109
Model:	GLM	Df Residuals:	1105
Model Family:	Binomial	Df Model:	3
Link Function:	logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-637.71
Date:	Thu, 09 Dec 2021	Deviance:	1275.4
Time:	15:02:35	Pearson chi2:	1.11e+03
No. Iterations:	4		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.3622	0.096	-14.119	0.000	-1.551	-1.173
Partisan Lean	-0.0150	0.004	-3.542	0.000	-0.023	-0.007
Unique Endorsements	-0.5497	0.265	-2.074	0.038	-1.069	-0.030
Proportion	16.4536	4.596	3.580	0.000	7.445	25.462

Figure 6: Frequentist

For our frequentist model we see we have a very negative log likelihood which may seem that this is not a good fit for our model. However if we look at the number of observations minus the number of parameters $n - p = 1109 - 5 = 1104$. Our deviance and chi-square χ^2 is 1275 and 1100 respectively because these are not too off our frequentist model does not seem to be a very good fit. Lets see our Bayesisan GLM using the binomial model. Note we chose binomial because of the link function is *logit*. We see that our model does converge which is quite good.

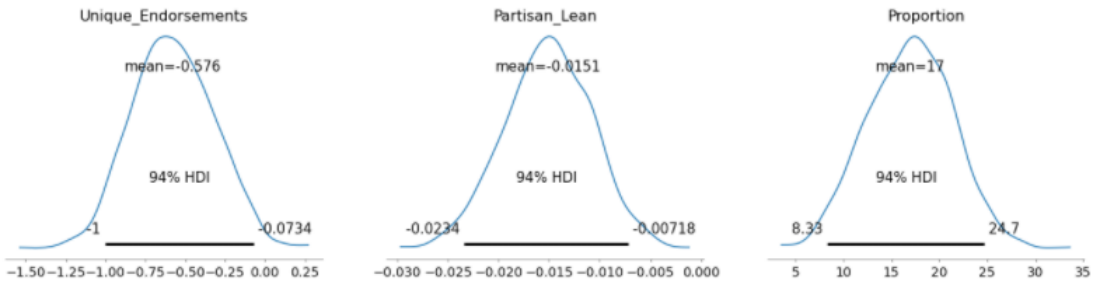


Figure 7: Bayesian Posterior

We seem to have an issue where our means are not very close to our true means. It seems that a Bayesian GLM doesn't do as well as the frequentist or the random forest and logistic regression.

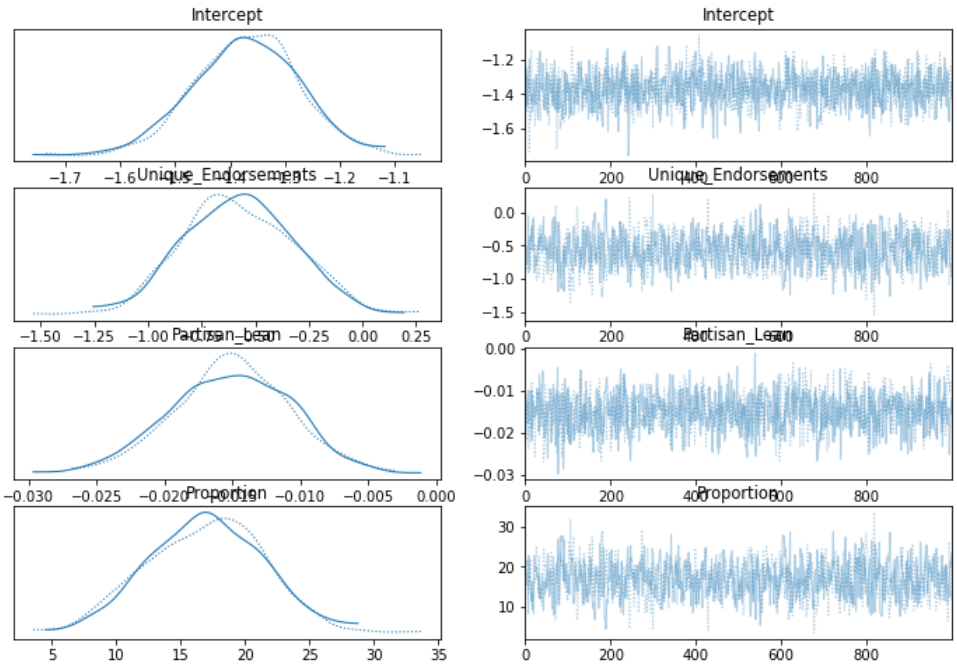


Figure 8: Bayesian

2.3 Discussion

Comparing to our nonparametric method of random forest we found we had a not terrible training accuracy but not very good test accuracy. We concluded this must be because of over fitting. However our GLM shows that indeed our model has some problems with our frequentist as well as our Bayesian model using posterior predictive checks. However the non-parametric method seems to be the bets. This model should not be used in future dataset. To improve the dataset it would be better if different data other than endorsements were added such as total contributions.

3 Bayesian Hierarchical Modeling

3.1 Methods

Based on our EDA we have an idea for our graphical model. We will try to estimate our parameter for adjusted leads. We observe first that $\mu_i \sim Uniform(-40, 40)$ adn that $\sigma^2 \sim HalfNormal(0, 5)$, then our likelihood function $X_i \mid (\mu, \sigma^2) \sim \mathcal{N}(\mu, \sigma^2)$. We will then introduce the regions of the U.S call them α_i whose value take on the set $\{0, 1, 2, 3\}$ where 0 corresponds to NE, 1 to MW, 2 to SO, 3 to WE.

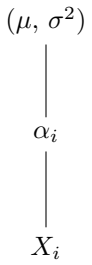
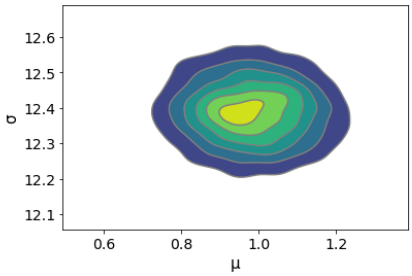


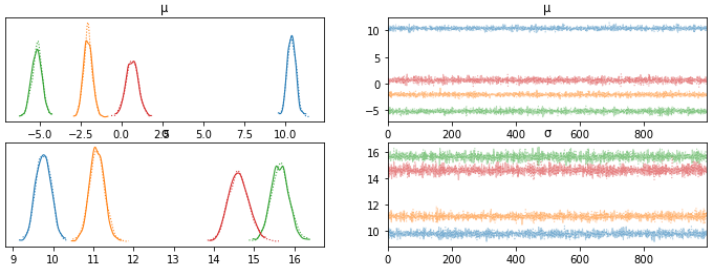
Figure 9: Graphical Model

3.2 Results

Our trace plots look good here. Our values on the left seem to have converged and be stationary and our MAP estimate which is the peak on the left graphs seem to be relatively close to the true value. Lets look at the joint trace of our sample. We can see here that our values do not seem to be correlated with each other which is a good thing. We can the individual influence the groups have on the adjusted leads between Trump and Clinton.



(a)



(b)

Figure 10: Trace plot with joint distribution

Lets evaluate our model. Running a posterior predictive check on all the regions of our model we see that our mean is very close to the true mean, denoted by the red line. Thus our model does a good job approximating our variable.

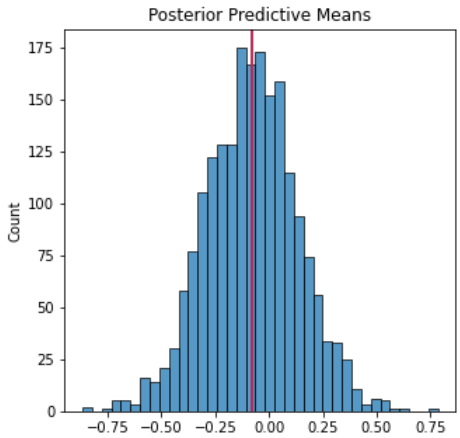


Figure 11: Posterior Predictive Check

3.3 Discussion

Our model is limited to the fact that its hard to choose priors for our model. We wanted to include poll weight in our model but couldn't find a prior that matched well with our model. This is one of the fundamental problems with Bayesian Inference. We didn't have trouble converging and had quite an easy time with converge with pymc3. Additional data that would be useful would be data that would be easy to plot and identify a prior distribution for which would make our model more robust.

4 Conclusion

We found that when evaluating endorsements on our Primary Elections datset random forest performed better than logistic regression. This is primarily due to the fact that our training set was probably hard to linearly seperate. However, random forest ran into issues with over-fitting. To avoid this perhaps we could have introduced regularization to tune our model complexity. These results are not very generalizable.

In our Bayesian Hierarchical Model, we found we were able to estimate adjusted leads between Trump and Clinton for the 2016 election quite well. Our model checks showed that our estimated mean was close to the true mean. Our model also converged well. This could be generalized a little to other datasets but would require tuning such as making sure the prior distributions for the new datasets aligned with the ones we chose.

Based on our findings, it is possible to predict elections given the right data. However, with polling data it only tells us sentiment rather than accurate predictions of the possible victor. For example most of the polling data points to Clinton winning but we all know how that turned out. Some limitations were overall domain knowledge pertaining to elections. Prior knowledge could have maybe led to choose a better GLM model leading to less error. Future studies could use our work go deeper into the implicit biases into polls and how they affect our elections.