Sample Size and Model Accuracy in Logistic Support Modeling

An examination of the relationship between sample size and logistic support model accuracy using data from the 2020 election cycle.

Abstract

Every year the United States Election cycle allows its citizens to elect their preferred federal, state, and local leaders. Surveys collect valuable data on these preferences which can then be used to train models. Although these models are important, obtaining a large sample size of voters can be expensive and time-consuming. To address this issue, we produced three Logistic Regression models based on 2020 Presidential, Senate, and Gubernatorial election cycle data to identify the optimal sample sizes needed to generate informative electoral predictions. The results show that the optimal sample sizes for each race vary depending on the scale of the election. This finding suggests that it is possible to produce high-performing models with a fraction of polled individuals relative to typical polling practices.

Introduction

This paper explores the propensity to vote for party candidates by producing models with demographically representative survey data. These data include 45,027 responses from voters across numerous demographic groups in the 2020 election cycle. The analysis is split into three modeling groups:

- Presidential Model (split by registered & non-registered states)
- Senate Model (Kentucky)
- Gubernatorial Model (Virginia)

The ultimate goal of this analysis is to assess the trade-offs between survey response and model performance. These findings will help optimize sample representation, survey responses needed (sample size), and cost. The Methodology section describes the measures taken to determine best fit models and validate our results.

Methodology

Our teams used logistic support models as the primary vehicle of analysis, which effectively provide predicted probabilities (rather than linear coefficients) for each of our binary dependent variables. These models provide coefficients that are easily communicated, scored, and compared.

For each modeling team, we began the initial process of cleaning data by removing unnecessary data. For example, “`support_senate`” would only be a relevant variable for the Senate modeling team. We also imputed median and means for any missing values, recoded categorical variables, and removed others to address issues of multicollinearity.

After the initial cleaning, each team began the process of feature selection using supervised L1 regression models to identify significant features. Upon manual review and adjustment, we imputed these features into Logistic Support Models to form predictions on training, testing, and full data separately. Lastly, each team retrieved the feature coefficients for each variable for future usage.

For the validation process, we utilized the coefficients retrieved from the previous step to compare the differences between the true average values grouped by demographic variables and the predicted values grouped by demographic variables.

Lastly, we have chosen to use the ROC charts to assess model performance. By deciding the predicted output from our models, we generated insights into how the "rank order" of our predicted values lines up with the actual rank order of responses.

Once an acceptable degree of efficacy was reached for each of our models, we began to explore variations in accuracy rates for different sample sizes. We elected to use a Monte Carlo Simulation method (more detailed code will be included in the next section). Monte Carlo Analysis allowed us to predict the accuracy for a variety of sample sizes. Thus, by simulating different levels of sample sizes and running the data over the logistic model, we are able to identify the most suitable sample size to be used for our analysis.

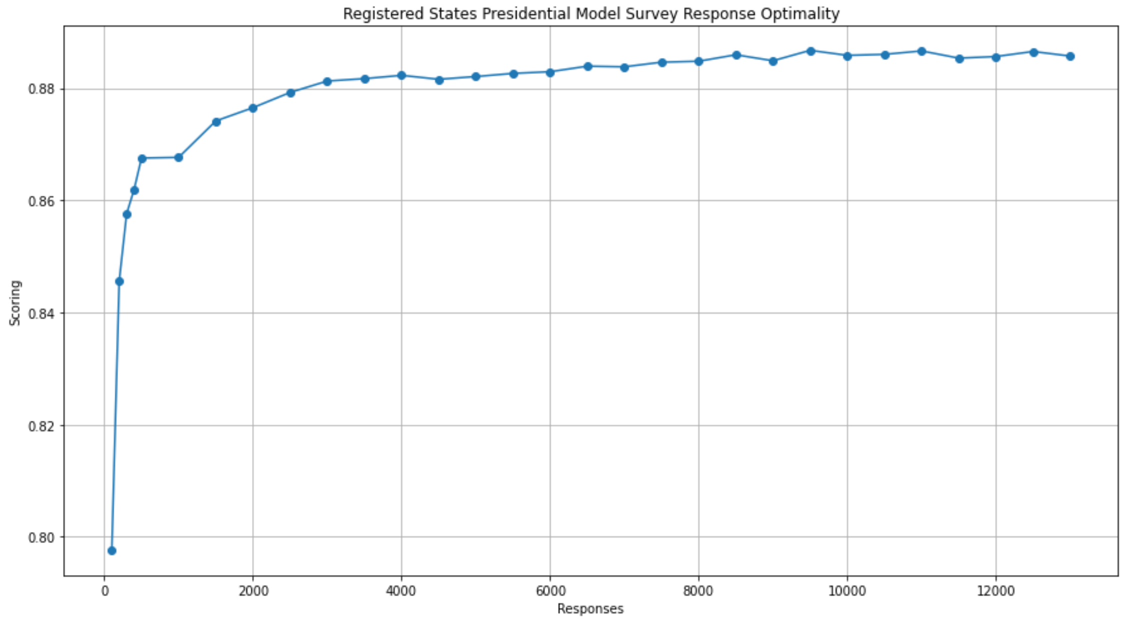
Pseudo-Code For Monte-Carlo Sample Estimation

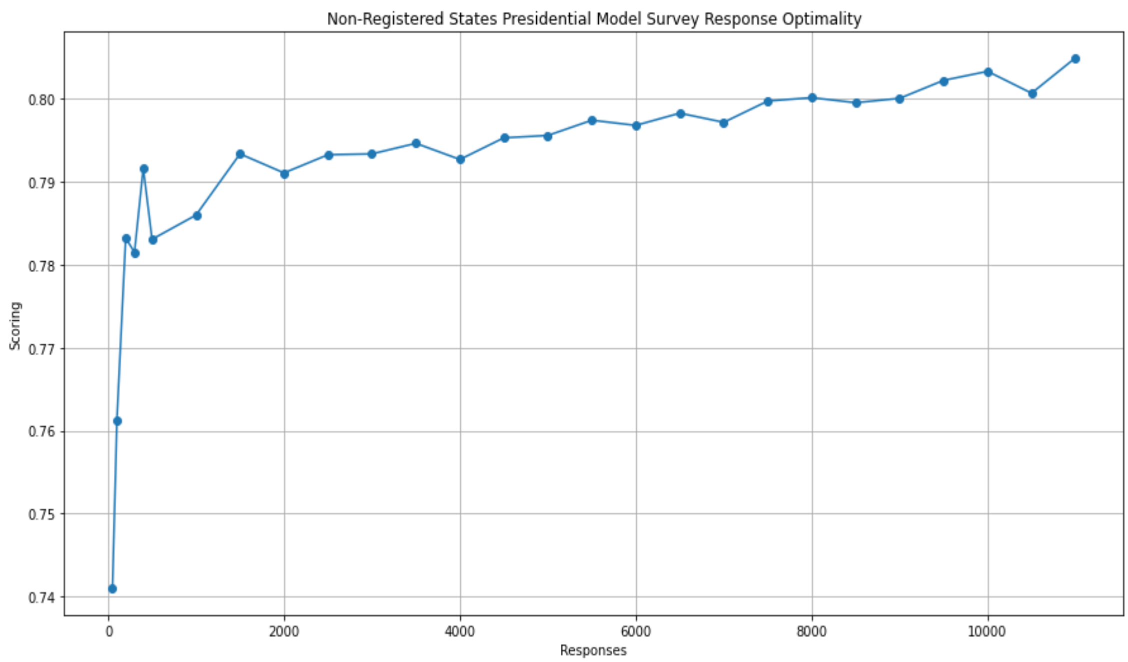
1. Define relevant series of sample sizes for each of the four models (presidential: registered states, presidential: non-registered states, gubernatorial, senate)
2. For each permutation of sample size:
    a. Generate pulls from the sample population using a stepwise function until maximum n size is reached
    b. For each subset of the sample from step a, control sample population using random seed from 1 to 100
    c. Run relevant logistic support model on each sample
    d. Generate training (75%) and testing (25%) data split
    e. Determine ROC scores for each training and testing sample permutation
3. Calculate summary statistics and plot cross-validation ROC score for each sample size permutation

Discussion & Findings

<u>Presidential</u>

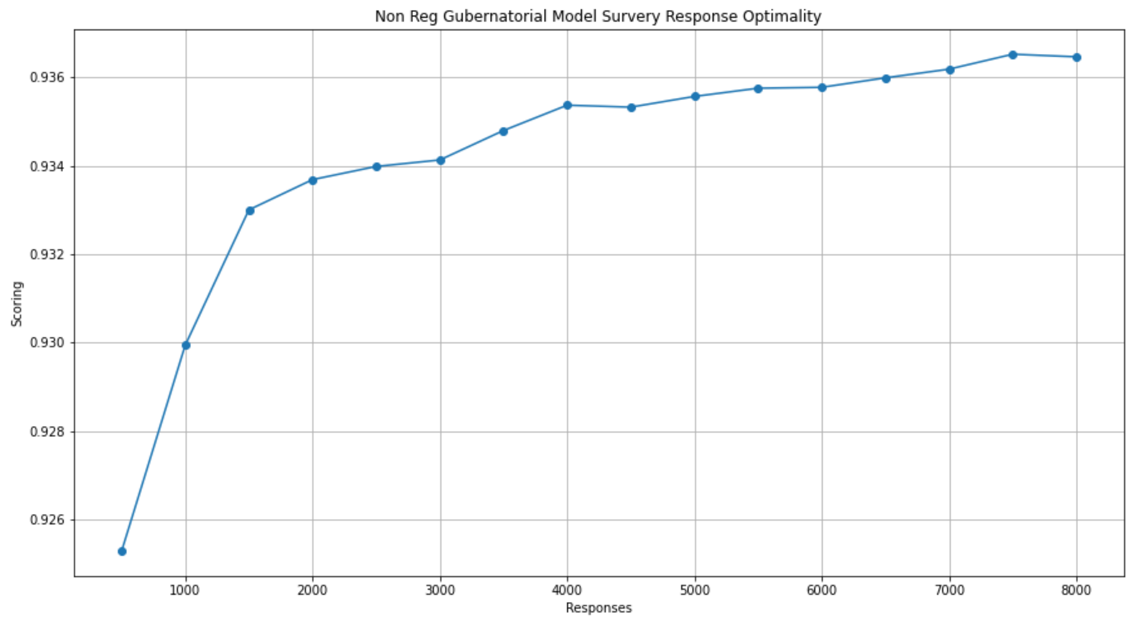
Below is the sample size chart for the Presidential Model. The accuracy rate for our registered states model tends to reach a horizontal asymptote at ~90%. The marginal utility of increasing n size begins to diminish significantly after a sample size of 2000, and an accuracy rate of 85% is possible with a sample as small as 1000. For our non-registered states model, the accuracy rates tend to be lower than registered states, with an upper limit of ~83%. A similar pattern of decreasing marginal utility at the 2000 mark is also apparent. An accuracy rate of 79% is possible with a sample of 2000 responses.





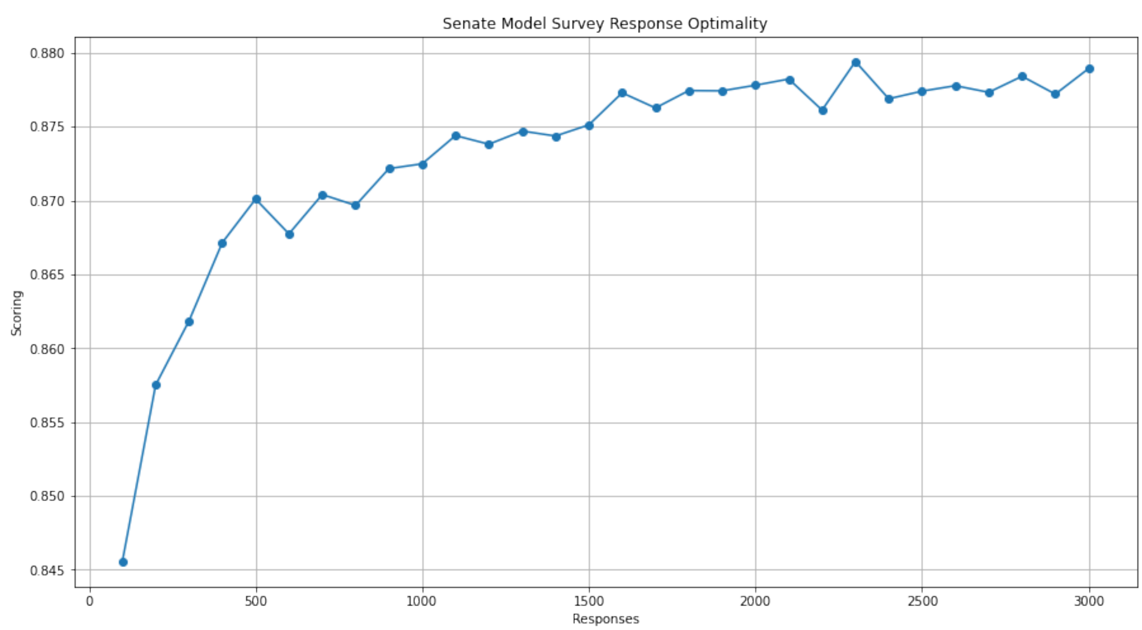
<u>Gubernatorial</u>

Below is the sample size chart for the Gubernatorial Model. An accuracy rate of near 88% is reached at an n size of roughly 2300. An accuracy of 87% possible with an n size of a small as 500.



<u>Senate</u>

Below is the sample size chart for the Senate Model. This model appeared to perform the best of the others at smaller sample sizes, with an accuracy of 93% possible with an n size of 1000. Similar to the other models, diminishing returns in accuracy are apparent after an n size of 4000.



Conclusion

In Graphs 1, 2, 3, and 4 we see the receiver operating characteristic curve, ROC curve, for each model. The ROC curve shows model performance across different thresholds. Graph 1 shows that the ROC curve for registered states in the Presidential model had an AUC of 0.947. Graph 2 shows that non-registered states in the Presidential model had an AUC of 0.901. Graph 4 shows the Gubernatorial model ROC curve which has an AUC of 0.932. Graph 3 shows the Senate model ROC curve which has an AUC of 0.822. The AUC values represent the accuracy of the models, the closer an AUC number is to 1, the more accurate the prediction. These differences in AUC_ROC scores suggest that model accuracies differ from sample subjects.

From the Monte Carlo Analysis, our results suggest that most of our models reach an upper-limit of accuracy between 80% and 90%, with the marginal utility of increasing n size beginning to diminish significantly after approximately 2000 samples. For all of our models other than the non-registered states presidential model, an accuracy score of 80+% was possible with an n size of 500. While these results suggest that we can still get decent accuracy rates with smaller (sometimes significantly so) sample sizes, one question for future research is whether the same pattern is repeated for accuracy at a more granular level. For example, good topline accuracy rates but perhaps less so for specific demographic groups or regions with models trained on smaller sample sizes.

For future research, other methods to explore include the Multilevel Regression and Poststratification (MRP) model. The MRP is a model that adjusts the results for known discrepancies between a sample population and a target population. This model would allow us to obtain better results for non-representative samples. For instance, an MRP would allow us to obtain more accurate results for states with small sample sizes.

Appendix 1: Results and Validations

Table 1: Presidential Model (Registered States)

Dep. Variable:	*support_pres*	*No. Observations:*	*9764*			
Model:	*Logit*	*Df Residuals:*	*9720*			
Method:	*MLE*	*Df Model:*	*43*			
Date:	*Wed*	*07 Dec 2022*	*Pseudo R-squ.:*	*0.5816*		
Time:	*0:04:58*	*Log-Likelihood:*	*-2816.5*			
converged:	*True*	*LL-Null:*	*-6732.1*			
Covariance Type:	*nonrobust*	*LLR p-value:*	*0*			
	coef	*std err*	*z*	*P>|z|*	*[0.025*	*0.975]*
const	*-2.0182*	*0.262*	*-7.706*	*0*	*-2.531*	*-1.505*
bInner Ring	*0.3137*	*0.076*	*4.118*	*0*	*0.164*	*0.463*
Income 125k+	*0.1346*	*0.089*	*1.51*	*0.131*	*-0.04*	*0.309*
age_a22to29	*0.7696*	*0.219*	*3.508*	*0*	*0.34*	*1.2*
age_b30to39	*0.6433*	*0.13*	*4.941*	*0*	*0.388*	*0.899*
black	*1.1664*	*0.168*	*6.923*	*0*	*0.836*	*1.497*
hispanic	*0.3293*	*0.167*	*1.967*	*0.049*	*0.001*	*0.657*

other_ethnicity	*0.6376*	*0.498*	*1.279*	*0.201*	*-0.339*	*1.614*
dem	*2.3836*	*0.093*	*25.679*	*0*	*2.202*	*2.565*
rep	*-2.5087*	*0.099*	*-25.365*	*0*	*-2.703*	*-2.315*
single	*0.6049*	*0.086*	*7.074*	*0*	*0.437*	*0.773*
homeowner	*0.0021*	*0.102*	*0.021*	*0.984*	*-0.199*	*0.203*
bach_degree_modeled	*0.3104*	*0.089*	*3.477*	*0.001*	*0.135*	*0.485*
post_grad_modeled	*0.348*	*0.103*	*3.389*	*0.001*	*0.147*	*0.549*
absentee_voter	*0.6332*	*0.081*	*7.821*	*0*	*0.475*	*0.792*
permanent_absentee_voter	*0.2076*	*0.158*	*1.314*	*0.189*	*-0.102*	*0.517*
midterm_primary_ever	*-0.0662*	*0.086*	*-0.771*	*0.441*	*-0.235*	*0.102*
indy_primary_only	*0.5506*	*0.41*	*1.343*	*0.179*	*-0.253*	*1.354*
dem_primary_mr	*1.2126*	*0.11*	*11.049*	*0*	*0.997*	*1.428*
switch_primary_dem_mr	*0.7319*	*0.304*	*2.409*	*0.016*	*0.136*	*1.327*
switch_primary_rep_mr	*-0.0874*	*0.207*	*-0.423*	*0.673*	*-0.493*	*0.318*
midterm_voter	*0.2482*	*0.132*	*1.874*	*0.061*	*-0.011*	*0.508*
non_voter	*0.5138*	*0.239*	*2.152*	*0.031*	*0.046*	*0.982*
rep_hhd	*1.6101*	*0.177*	*9.072*	*0*	*1.262*	*1.958*
indy_hhod	*1.4744*	*0.182*	*8.112*	*0*	*1.118*	*1.831*

hh_liberal_donor_u	*0.3442*	*0.236*	*1.456*	*0.145*	*-0.119*	*0.807*
christian	*-0.0025*	*0.09*	*-0.028*	*0.978*	*-0.179*	*0.174*
liberal_donor	*0.2421*	*0.106*	*2.286*	*0.022*	*0.034*	*0.45*
contbpol_1	*0.0732*	*0.109*	*0.673*	*0.501*	*-0.14*	*0.286*
contbhlt_1	*0.0088*	*0.102*	*0.086*	*0.931*	*-0.191*	*0.208*
professional_technical	*0.1422*	*0.115*	*1.234*	*0.217*	*-0.084*	*0.368*
apparel_1	*0.0699*	*0.09*	*0.775*	*0.439*	*-0.107*	*0.247*
outdgrdn_1	*0.1913*	*0.111*	*1.731*	*0.084*	*-0.025*	*0.408*
golf_1	*-0.139*	*0.136*	*-1.023*	*0.306*	*-0.405*	*0.127*
cnty_pct_religious	*0.24*	*0.329*	*0.73*	*0.465*	*-0.404*	*0.884*
evan_0to70	*0.174*	*0.137*	*1.266*	*0.205*	*-0.095*	*0.443*
evan_100to140	*0.0727*	*0.127*	*0.574*	*0.566*	*-0.176*	*0.321*
evan_140to230	*0.1382*	*0.124*	*1.112*	*0.266*	*-0.105*	*0.382*
evan_230plus	*0.2166*	*0.138*	*1.574*	*0.116*	*-0.053*	*0.486*
vote_pp2008_not_2012	*0.3489*	*0.11*	*3.173*	*0.002*	*0.133*	*0.564*
vote_g2012_or_g2008	*0.0458*	*0.129*	*0.356*	*0.722*	*-0.207*	*0.298*
vote_g2006_not_2010	*0.2231*	*0.157*	*1.421*	*0.155*	*-0.085*	*0.531*
early_vote_method_g2018	*-0.0718*	*0.159*	*-0.452*	*0.651*	*-0.383*	*0.239*

ineligible_2008to2022	*0.2133*	*0.115*	*1.85*	*0.064*	*-0.013*	*0.439*

Table 2: Presidential Model (Non-Registered States)

Dep. Variable:	*support_pres*	*No. Observations:*	*8308*			

Model:	*Logit*	*Df Residuals:*	*8266*			
Method:	*MLE*	*Df Model:*	*41*			
Date:	*Wed*	*07 Dec 2022*	*Pseudo R-squ.:*	*0.4644*		
Time:	*0:03:11*	*Log-Likelihood:*	*-3084.1*			
converged:	*True*	*LL-Null:*	*-5758.7*			
Covariance Type:	*nonrobust*	*LLR p-value:*	*0*			
	coef	*std err*	*z*	*P>|z|*	*[0.025*	*0.975]*
const	*-3.7289*	*0.216*	*-17.229*	*0*	*-4.153*	*-3.305*
age_a22to29	*0.5776*	*0.203*	*2.851*	*0.004*	*0.18*	*0.975*
age_b30to39	*0.8516*	*0.123*	*6.948*	*0*	*0.611*	*1.092*
age_c40to49	*0.6015*	*0.093*	*6.494*	*0*	*0.42*	*0.783*
black	*1.1679*	*0.125*	*9.33*	*0*	*0.923*	*1.413*
hispanic	*0.295*	*0.174*	*1.697*	*0.09*	*-0.046*	*0.636*
asian	*0.8563*	*0.373*	*2.299*	*0.022*	*0.126*	*1.586*
other_ethnicity	*1.7241*	*0.513*	*3.358*	*0.001*	*0.718*	*2.731*

single	*0.5672*	*0.08*	*7.06*	*0*	*0.41*	*0.725*
length_of_residence	*0.0054*	*0.007*	*0.766*	*0.444*	*-0.008*	*0.019*
children	*0.0008*	*0.081*	*0.01*	*0.992*	*-0.157*	*0.159*
homeowner	*-0.1225*	*0.105*	*-1.17*	*0.242*	*-0.328*	*0.083*
bach_degree_modeled	*0.2282*	*0.078*	*2.925*	*0.003*	*0.075*	*0.381*
post_grad_modeled	*0.142*	*0.094*	*1.513*	*0.13*	*-0.042*	*0.326*
inactive_voter	*0.3574*	*0.173*	*2.063*	*0.039*	*0.018*	*0.697*
absentee_voter	*1.1134*	*0.07*	*15.797*	*0*	*0.975*	*1.251*
permanent_absentee_voter	*-0.1051*	*0.132*	*-0.794*	*0.427*	*-0.364*	*0.154*
midterm_primary_ever	*-0.44*	*0.081*	*-5.453*	*0*	*-0.598*	*-0.282*
indy_primary_only	*3.046*	*1.25*	*2.437*	*0.015*	*0.597*	*5.495*
dem_primary_mr	*4.7559*	*0.167*	*28.535*	*0*	*4.429*	*5.083*
switch_primary_dem_mr	*-1.6268*	*0.175*	*-9.282*	*0*	*-1.97*	*-1.283*
hh_black_u	*0.2326*	*0.258*	*0.903*	*0.367*	*-0.272*	*0.738*
hh_liberal_donor_u	*0.4751*	*0.215*	*2.207*	*0.027*	*0.053*	*0.897*
hasreligion	*0.0974*	*0.087*	*1.113*	*0.266*	*-0.074*	*0.269*
liberal_donor	*0.428*	*0.089*	*4.796*	*0*	*0.253*	*0.603*
contbrel_1	*-0.0274*	*0.099*	*-0.277*	*0.782*	*-0.222*	*0.167*

contbpol_1	*-0.0465*	*0.115*	*-0.403*	*0.687*	*-0.273*	*0.18*
blue_collar	*0.1321*	*0.112*	*1.183*	*0.237*	*-0.087*	*0.351*
professional_technical	*0.2086*	*0.103*	*2.02*	*0.043*	*0.006*	*0.411*
apparel_1	*0.0775*	*0.083*	*0.935*	*0.35*	*-0.085*	*0.24*
outdgrdn_1	*0.0383*	*0.096*	*0.399*	*0.69*	*-0.15*	*0.226*
expensive_items_1	*-0.0057*	*0.086*	*-0.066*	*0.947*	*-0.174*	*0.163*
cnty_pct_religious	*3.3594*	*0.3*	*11.202*	*0*	*2.772*	*3.947*
evan_0to70	*1.0799*	*0.167*	*6.459*	*0*	*0.752*	*1.408*
evan_100to140	*0.2558*	*0.084*	*3.032*	*0.002*	*0.09*	*0.421*
evan_230plus	*-0.7495*	*0.083*	*-8.983*	*0*	*-0.913*	*-0.586*
vote_pp2008_not_2012	*0.5312*	*0.109*	*4.858*	*0*	*0.317*	*0.746*
vote_g2006_not_2010	*0.1785*	*0.133*	*1.337*	*0.181*	*-0.083*	*0.44*
vote_pp_or_p2008	*-0.2121*	*0.089*	*-2.381*	*0.017*	*-0.387*	*-0.037*
greatest_year_d	*2.04E-05*	*5.99E-05*	*0.341*	*0.733*	*-9.70E-05*	*0*
greatest_year_i	*9.02E-05*	*0*	*0.388*	*0.698*	*0*	*0.001*
ineligible_2008to2022	*0.0854*	*0.097*	*0.877*	*0.381*	*-0.106*	*0.276*

Table 3: Senate Model

Dep. Variable:	*support_senate*	*No. Observations:*	*2322*			
Model:	*Logit*	*Df Residuals:*	*2271*			
Method:	*MLE*	*Df Model:*	*50*			
Date:	*Tue, 13 Dec 2022*	*Pseudo R-squ.:*	*0.4107*			
Time:	*5:48:44*	*Log-Likelihood:*	*-937.16*			
converged:	*TRUE*	*LL-Null:*	*-1590.3*			

Covariance Type:	*nonrobust*	*LLR p-value:*	*1.33E-240*			
coef	*std err*	*z*	*P>|z|*	*[0.025*	*0.975]*	
const	*-0.0234*	*3.82E+06*	*-6.14E-09*	*1*	*-7.49E+06*	*7.49E+06*
bInner Ring	*-0.4618*	*0.409*	*-1.129*	*0.259*	*-1.264*	*0.34*
cOuter Ring	*-1.0552*	*0.405*	*-2.603*	*0.009*	*-1.85*	*-0.261*
dRural	*-1.3553*	*0.406*	*-3.342*	*0.001*	*-2.15*	*-0.56*
Income 030k-50k	*0.1168*	*0.211*	*0.554*	*0.58*	*-0.297*	*0.53*
Income 050k-75k	*-0.2176*	*0.199*	*-1.096*	*0.273*	*-0.607*	*0.171*
Income 075k-125k	*-0.0911*	*0.201*	*-0.453*	*0.65*	*-0.485*	*0.303*
Income 125k+	*0.2194*	*0.225*	*0.976*	*0.329*	*-0.221*	*0.66*
age_a22to29	*1.2051*	*0.406*	*2.966*	*0.003*	*0.409*	*2.001*
age_b30to39	*0.5759*	*0.258*	*2.236*	*0.025*	*0.071*	*1.081*
age_c40to49	*0.4082*	*0.217*	*1.879*	*0.06*	*-0.018*	*0.834*
age_g80plus	*0.1423*	*0.231*	*0.616*	*0.538*	*-0.311*	*0.595*
gender_male	*-0.4002*	*0.121*	*-3.295*	*0.001*	*-0.638*	*-0.162*
black	*0.8571*	*0.254*	*3.377*	*0.001*	*0.36*	*1.355*
hispanic	*3.5368*	*1.34E+00*	*2.632*	*0.008*	*9.03E-01*	*6.17E+00*
dem	*0.9838*	*0.26*	*3.78*	*0*	*0.474*	*1.494*
rep	*-2.5965*	*0.278*	*-9.344*	*0*	*-3.141*	*-2.052*
single	*0.682*	*0.15*	*4.541*	*0*	*0.388*	*0.976*
renter	*0.6934*	*0.464*	*1.496*	*0.135*	*-0.215*	*1.602*
bach_degree_modeled	*0.3966*	*0.201*	*1.973*	*0.048*	*0.003*	*0.791*
post_grad_modeled	*0.3036*	*0.196*	*1.547*	*0.122*	*-0.081*	*0.688*
absentee_voter	*0.3985*	*0.367*	*1.086*	*0.277*	*-0.321*	*1.118*
midterm_primary_ever	*0.331*	*0.158*	*2.092*	*0.036*	*0.021*	*0.641*
midterm_voter	*-0.2842*	*3.45E+06*	*-8.25E-08*	*1*	*-6.75E+06*	*6.75E+06*
presidential_voter	*-0.2779*	*3.45E+06*	*-8.07E-08*	*1*	*-6.75E+06*	*6.75E+06*

new_registrant	*0.6395*	*3.45E+06*	*1.86E-07*	*1*	*-6.75E+06*	*6.75E+06*
non_voter	*-0.1008*	*3.45E+06*	*-2.93E-08*	*1*	*-6.75E+06*	*6.75E+06*
rep_hhd	*0.84*	*0.291*	*2.885*	*0.004*	*0.269*	*1.411*
indy_hhod	*0.6612*	*0.489*	*1.351*	*0.177*	*-0.298*	*1.62*
hh_liberal_donor_u	*0.1559*	*0.39*	*0.399*	*0.69*	*-0.609*	*0.921*
hasreligion	*-0.0131*	*0.143*	*-0.092*	*0.927*	*-0.293*	*0.267*
liberal_donor	*-0.0855*	*0.19*	*-0.45*	*0.653*	*-0.458*	*0.287*
contbpol_1	*0.4222*	*0.218*	*1.94*	*0.052*	*-0.004*	*0.849*
contbhlt_1	*0.1129*	*0.174*	*0.649*	*0.516*	*-0.228*	*0.454*
professional_technical	*0.4081*	*0.198*	*2.062*	*0.039*	*0.02*	*0.796*
retired	*0.1881*	*0.338*	*0.556*	*0.578*	*-0.475*	*0.851*
bookmusc_1	*-0.1289*	*0.185*	*-0.699*	*0.485*	*-0.491*	*0.233*
outdgrdn_1	*0.0794*	*0.183*	*0.434*	*0.664*	*-0.279*	*0.438*
expensive_items_1	*0.1071*	*0.172*	*0.622*	*0.534*	*-0.23*	*0.444*
evan_140to230	*0.6235*	*0.566*	*1.101*	*0.271*	*-0.486*	*1.733*
evan_230plus	*0.4631*	*0.548*	*0.845*	*0.398*	*-0.611*	*1.538*
vote_g2012_or_g2008	*-0.0421*	*0.216*	*-0.195*	*0.845*	*-0.465*	*0.381*
vote_g2006_not_2010	*0.548*	*0.232*	*2.365*	*0.018*	*0.094*	*1.002*
vote_pp_or_p2008	*0.3434*	*0.141*	*2.439*	*0.015*	*0.067*	*0.619*
ineligible_2008to2022	*0.1767*	*0.235*	*0.753*	*0.451*	*-0.283*	*0.637*
aUrban	*-0.5735*	*0.305*	*-1.878*	*0.06*	*-1.172*	*0.025*
hs_or_less	*0.2301*	*0.133*	*1.736*	*0.083*	*-0.03*	*0.49*
low_edu_area	*-0.0065*	*3.71E+06*	*-1.74E-09*	*1*	*-7.28E+06*	*7.28E+06*
med_edu_area	*-0.0609*	*3.71E+06*	*-1.64E-08*	*1*	*-7.28E+06*	*7.28E+06*
high_edu_area	*0.044*	*3.71E+06*	*1.18E-08*	*1*	*-7.28E+06*	*7.28E+06*
age_d50to64	*-0.002*	*0.197*	*-0.01*	*0.992*	*-0.388*	*0.384*
age_e65plus	*-0.0334*	*0.186*	*-0.18*	*0.857*	*-0.398*	*0.331*

independents	*-0.3964*	*0.351*	*-1.128*	*0.259*	*-1.085*	*0.292*
independents	*0.1592*	*0.434*	*0.367*	*0.714*	*-0.691*	*1.01*

Table 4: Gubernatorial Model

Dep. Variable:	*support_gov*	*No. Observations:*	*6014*			

Model:	Logit	Df Residuals:	5978			
Method:	MLE	Df Model:	35			
Date:	Wed	14 Dec 2022	Pseudo R-squared:	0.5195		
Time:	19:24:19	Log-Likelihood:	-2000.7			
converged:	False	LL-Null:	-4163.8			
Covariance Type:	nonrobust	LLR p-value:	0			
	coef	std err	z	P>|z|	[0.025	0.975]
const	1.4889	1.091	1.365	0.172	-0.649	3.627
Income 075k-125k	0.0608	0.168	0.363	0.717	-0.268	0.389
vote_g2010_not_2006	-0.3077	0.127	-2.419	0.016	-0.557	-0.058
age_b30to39	-2.4235	1.083	-2.237	0.025	-4.547	-0.3
Income 030k-50k	0.1679	0.185	0.907	0.364	-0.195	0.531
conservative_donor	-1.9869	0.329	-6.043	0	-2.631	-1.342
greatest_year_d	0.001	5.86E-05	17.835	0	0.001	0.001
rep_primary_mr	-2.2713	0.153	-14.81	0	-2.572	-1.971
age_c40to49	-2.5056	1.083	-2.314	0.021	-4.628	-0.384
asian	1.3399	0.355	3.771	0	0.644	2.036
age_e60to69	-2.5665	1.082	-2.373	0.018	-4.687	-0.446
single	0.4464	0.092	4.835	0	0.265	0.627
bach_degree_modeled	0.0243	0.108	0.224	0.823	-0.188	0.236
bInner Ring	-0.1041	0.137	-0.757	0.449	-0.374	0.165
other_ethnicity	18.8082	1516.581	0.012	0.99	-2953.636	2991.253

early_voter	*-0.287*	*0.205*	*-1.402*	*0.161*	*-0.688*	*0.114*
absentee_voter	*1.0524*	*0.103*	*10.178*	*0*	*0.85*	*1.255*
vote_pp2008_not_2012	*0.4639*	*0.121*	*3.826*	*0*	*0.226*	*0.702*
age_d50to59	*-2.5858*	*1.081*	*-2.391*	*0.017*	*-4.705*	*-0.466*
age_g80plus	*-2.291*	*1.096*	*-2.091*	*0.037*	*-4.439*	*-0.143*
Income 050k-75k	*-0.0609*	*0.172*	*-0.353*	*0.724*	*-0.399*	*0.277*
age_f70to79	*-2.4572*	*1.084*	*-2.268*	*0.023*	*-4.581*	*-0.333*
vote_pp_or_p2012	*-0.811*	*0.13*	*-6.228*	*0*	*-1.066*	*-0.556*
gender_male	*-0.3638*	*0.081*	*-4.512*	*0*	*-0.522*	*-0.206*
black	*0.9478*	*0.122*	*7.749*	*0*	*0.708*	*1.188*
hispanic	*0.5573*	*0.275*	*2.029*	*0.042*	*0.019*	*1.096*
Income 125k+	*0.0844*	*0.167*	*0.504*	*0.614*	*-0.244*	*0.413*
cOuter Ring	*-0.539*	*0.139*	*-3.876*	*0*	*-0.812*	*-0.266*
dRural	*-0.7637*	*0.153*	*-4.986*	*0*	*-1.064*	*-0.463*
retired	*-0.3378*	*0.283*	*-1.195*	*0.232*	*-0.892*	*0.216*
age_a22to29	*-2.306*	*1.096*	*-2.105*	*0.035*	*-4.453*	*-0.159*
liberal_donor	*3.0964*	*0.63*	*4.918*	*0*	*1.862*	*4.33*
environm_1	*-2.9524*	*0.631*	*-4.678*	*0*	*-4.189*	*-1.716*
midterm_primary_ever	*0.546*	*0.114*	*4.778*	*0*	*0.322*	*0.77*
greatest_year_r	*-0.0007*	*5.83E-05*	*-12.78*	*0*	*-0.001*	*-0.001*
post_grad_modeled	*0.2336*	*0.112*	*2.085*	*0.037*	*0.014*	*0.453*

Appendix 2: Decile Charts

Chart 1: Presidential Model (Registered States) Decile Chart

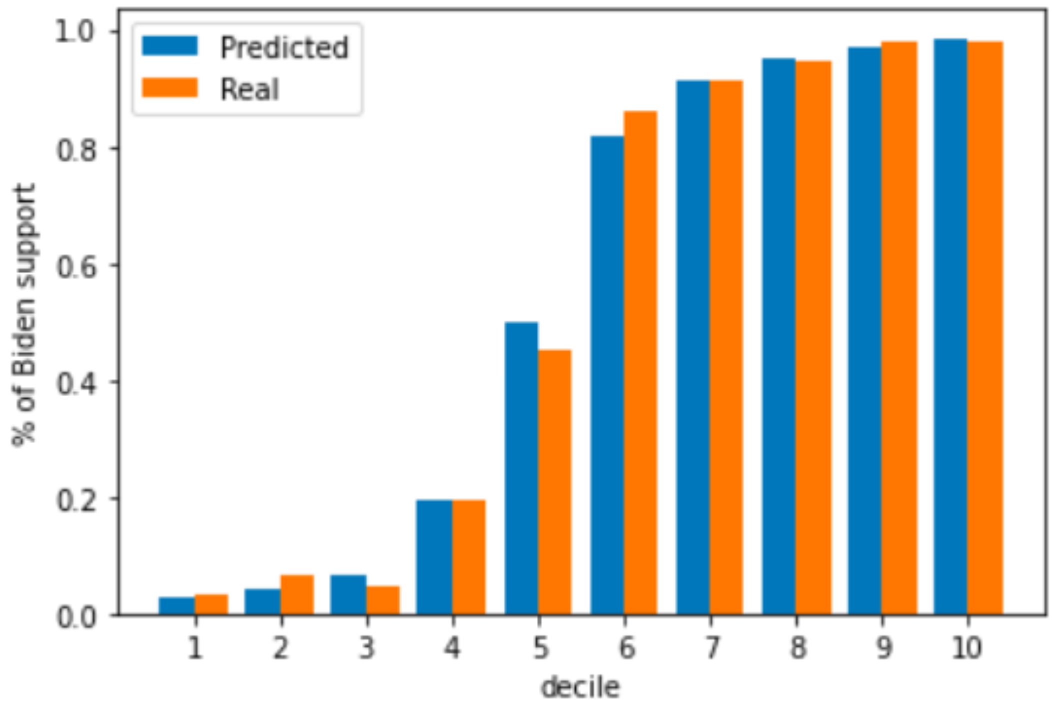


Chart 2: Presidential Model (Non-Registered States) Decile Chart

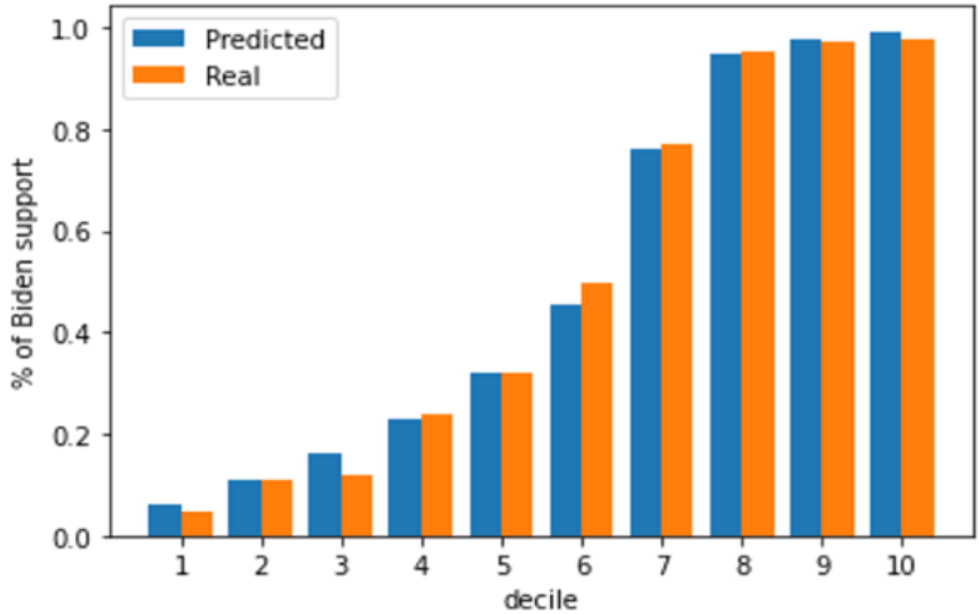


Chart 3: Senate Model Decile Chart

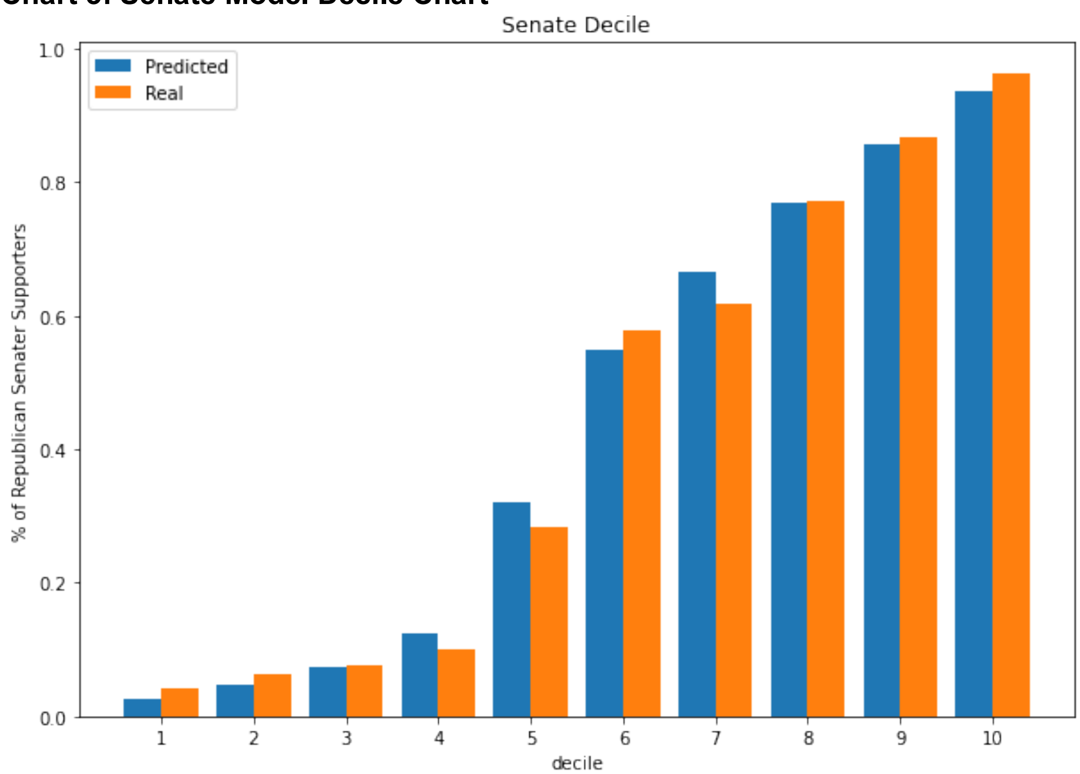
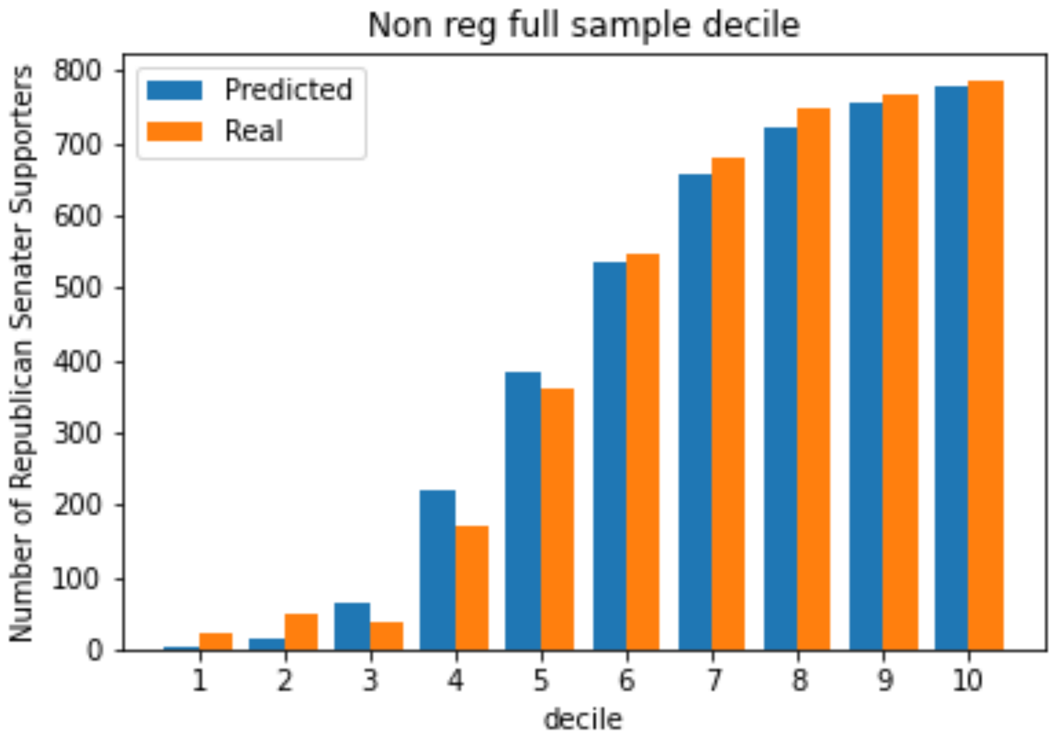
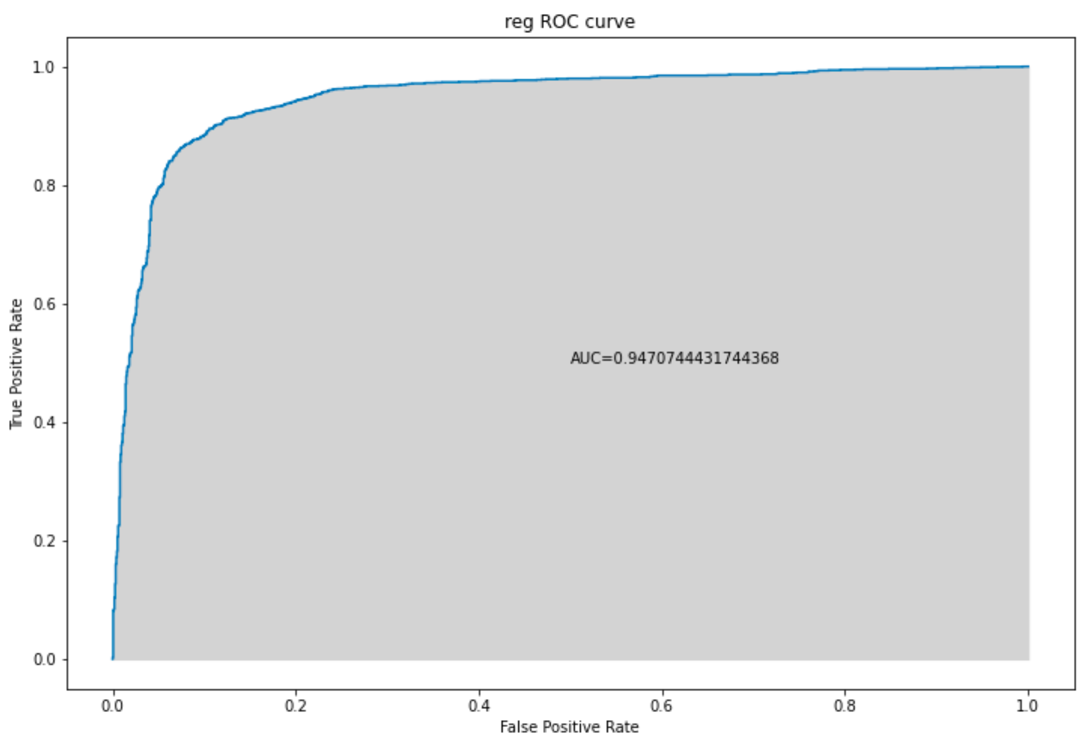


Chart 4: Governor Model Decile Chart

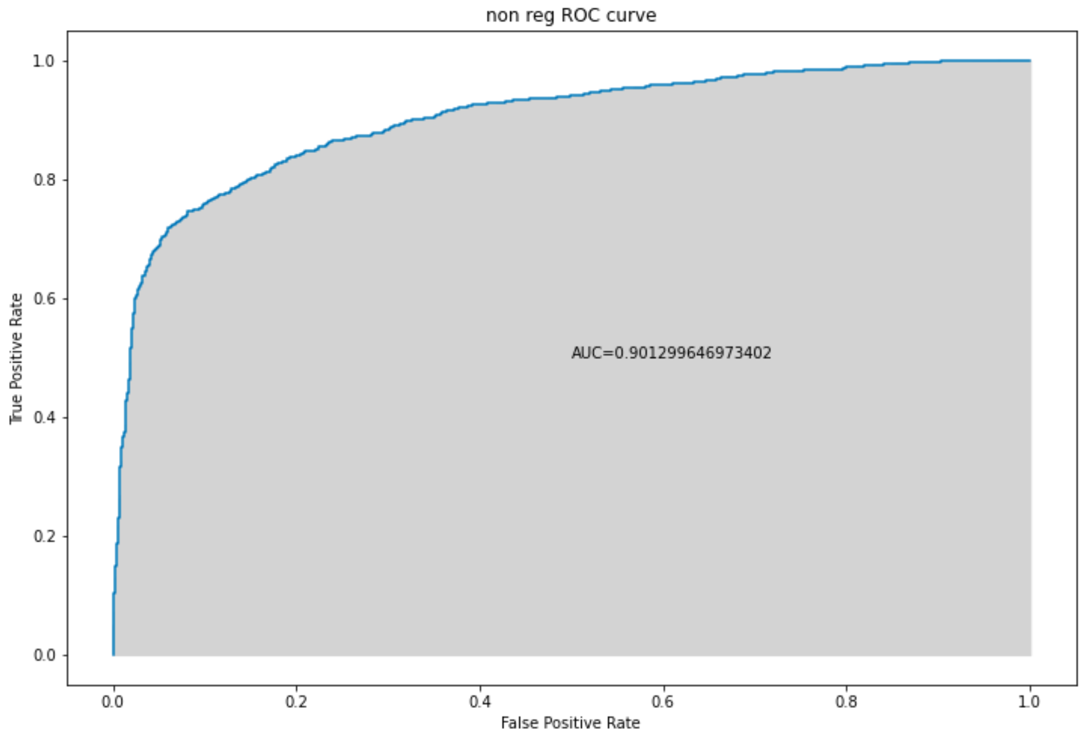


Appendix 3: ROC Curves

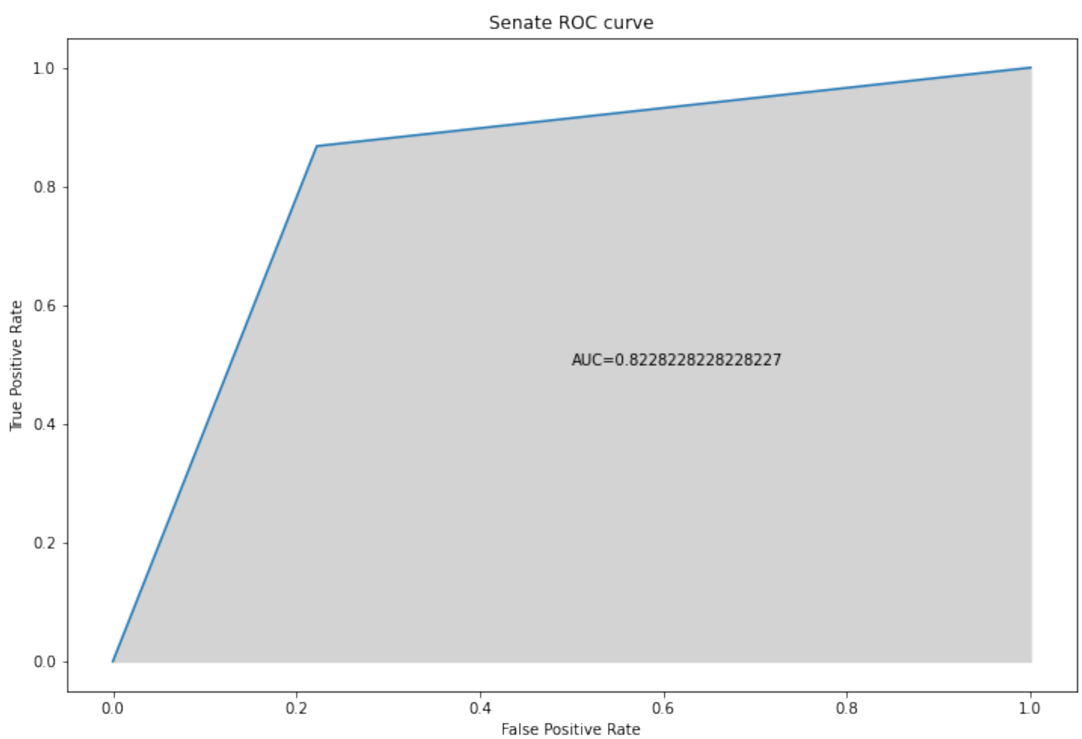
Graph 1: Presidential Model (Registered States) ROC Curve



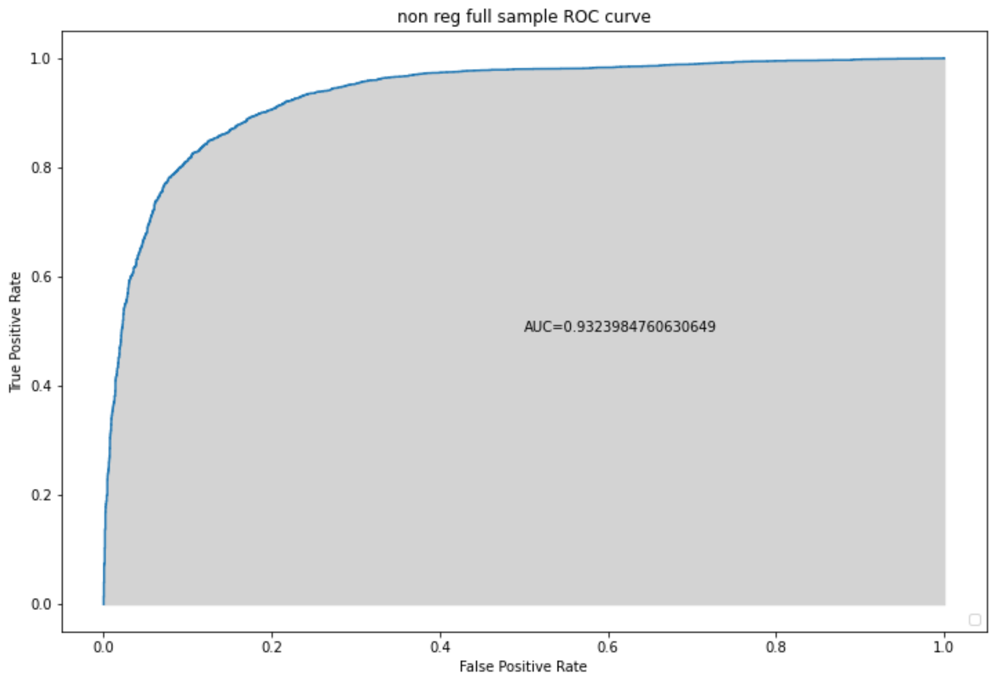
Graph 2: Presidential Model (Non-Registered States) ROC Curve



Graph 3: Senate Model ROC Curve



Graph 4: Governor Model ROC Curve


non reg full sample ROC curve
AUC=0.9323984760630649
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0