Predicting Voter Turnout In North Carolina

Phillip Dibert

General Assembly Data Science Immersive

DSI-7-NY-Nash

Table of Contents

Problem Statement

- Using the North Carolina voter file (voter history and demographics), can we use data from 2010 - 2014 to predict whether one million randomly selected voters voted in the 2016 general election?

- Will adding in congressional district level donation data improve our models?

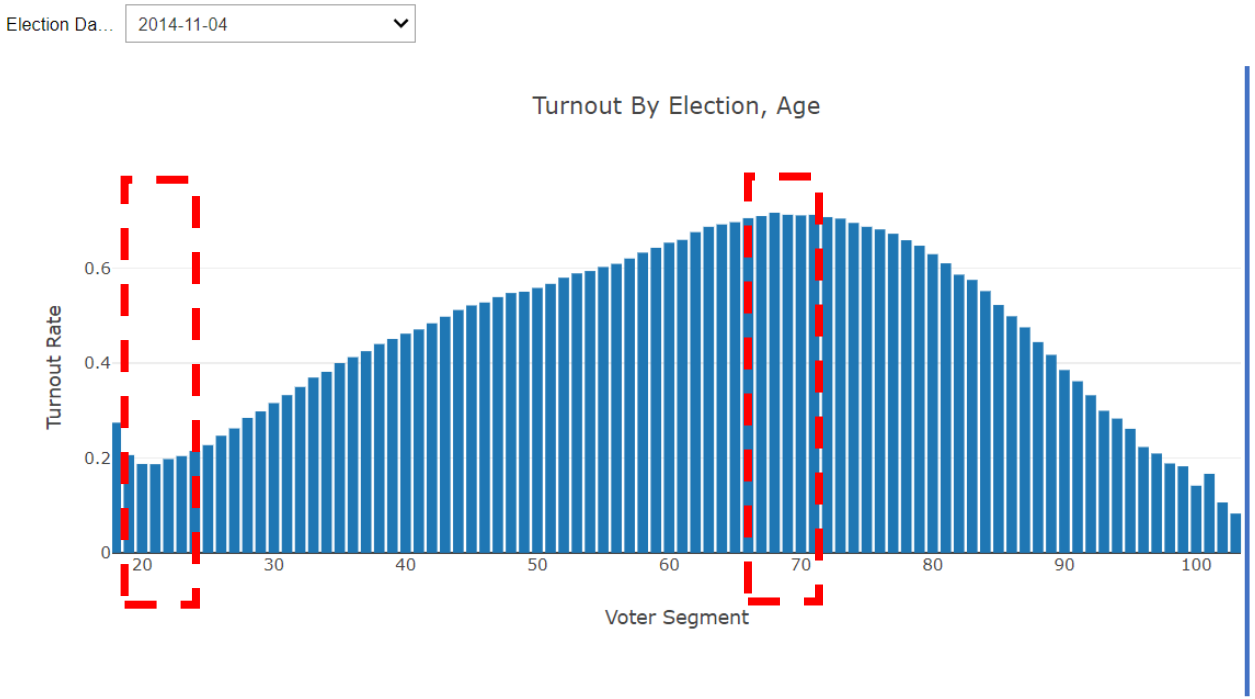
Background Information

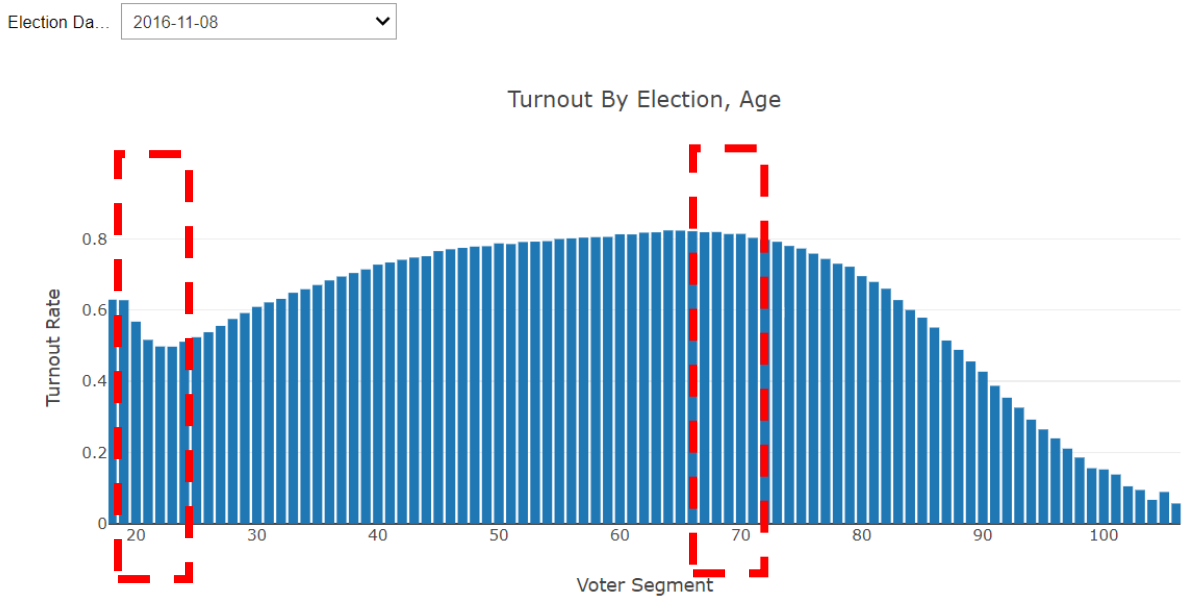
- North Carolina law mandates that voter information be made available to the public
- Political campaigns use this information to identify likely voters so that they can communicate with voters, identify strategies, and plan media spend
- Voter file can be joined to other datasets to give a fuller picture of the political landscape

Exploratory Data Visualizations

Voter participation and donations

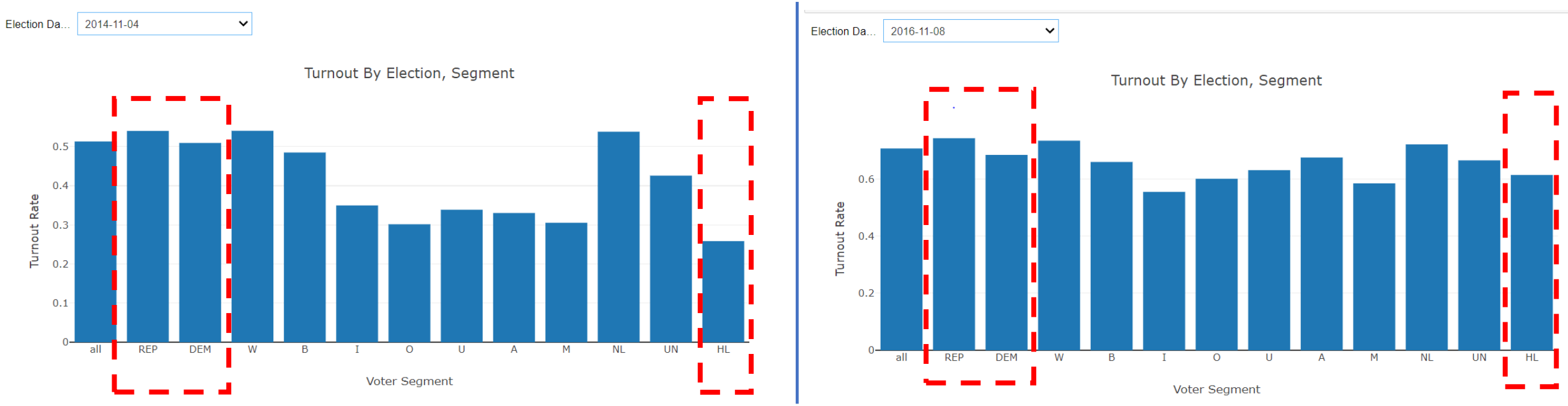
Young people turnout at much lower rates than older voters





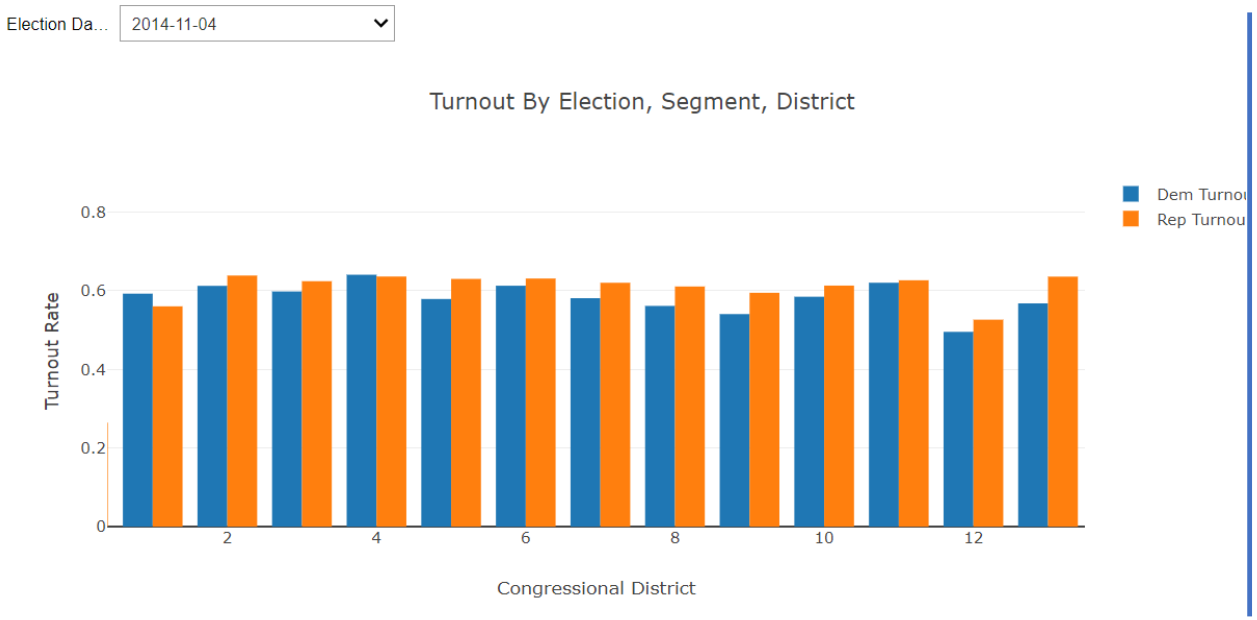
- In 2014, voters aged 67-71 turned out at rates of 71%, while those 20-24 had rates of 20% and below
- In 2016, the turnout gap was 30% between those ages 67-71(~80%) and 20-24(~50%)

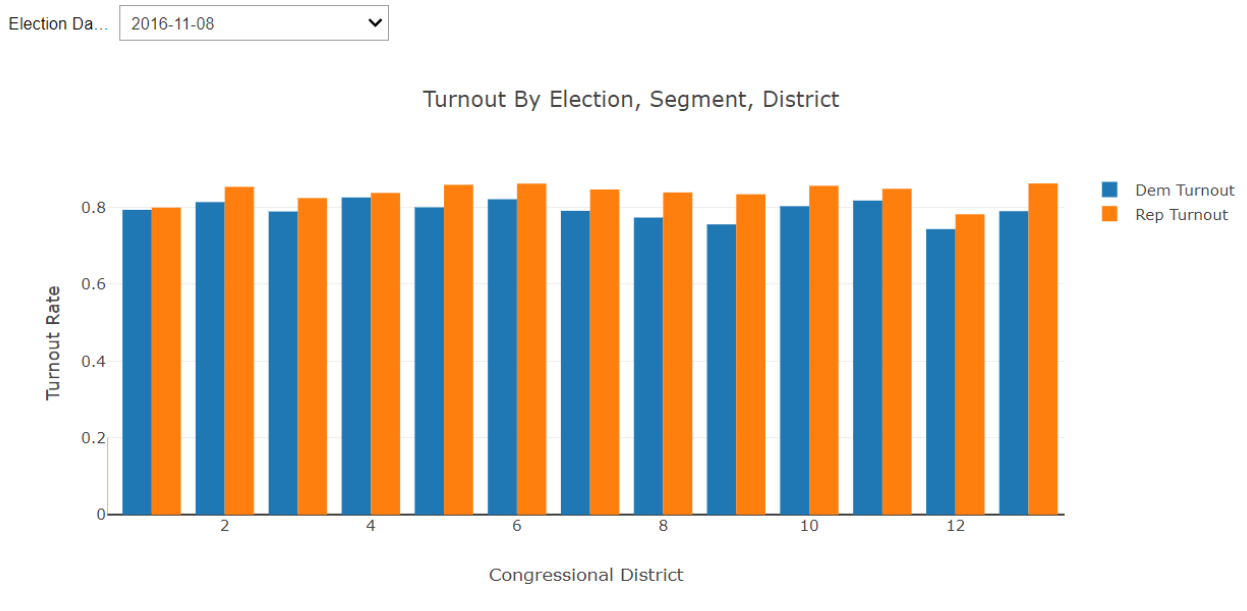
Turnout By Demographic Segment (2014, 2016)



- Turnout among Republicans(2014: 54%, 2016: 74%) was higher relative to that of Democrats(51%, 68%) in both 2014 and 2016
- The sharpest increase between the 2014 and 2016 cycles occurred among Latinos (+137% CoC)

Republican turnout exceeded that of Democrats for all congressional districts in 2016

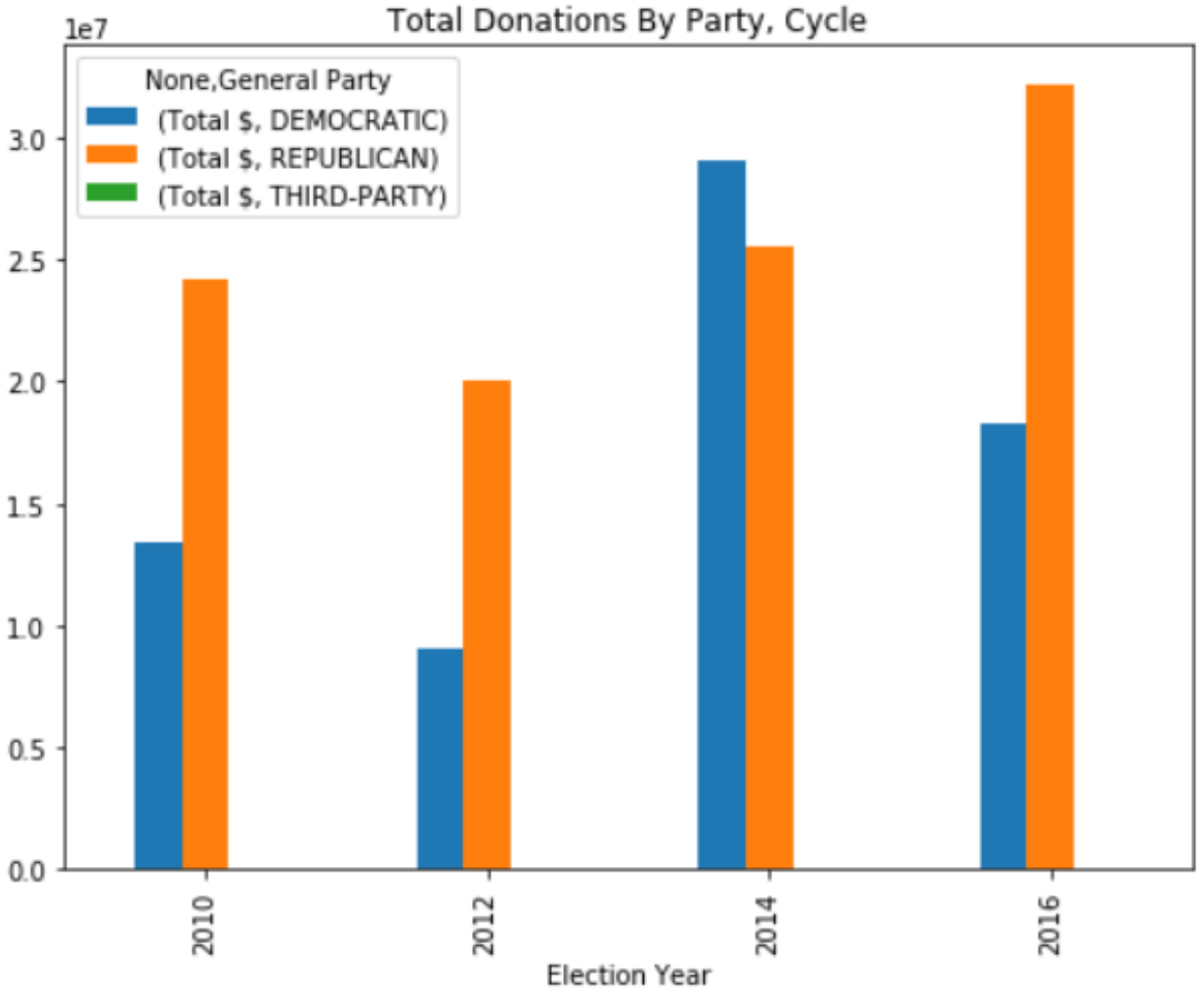




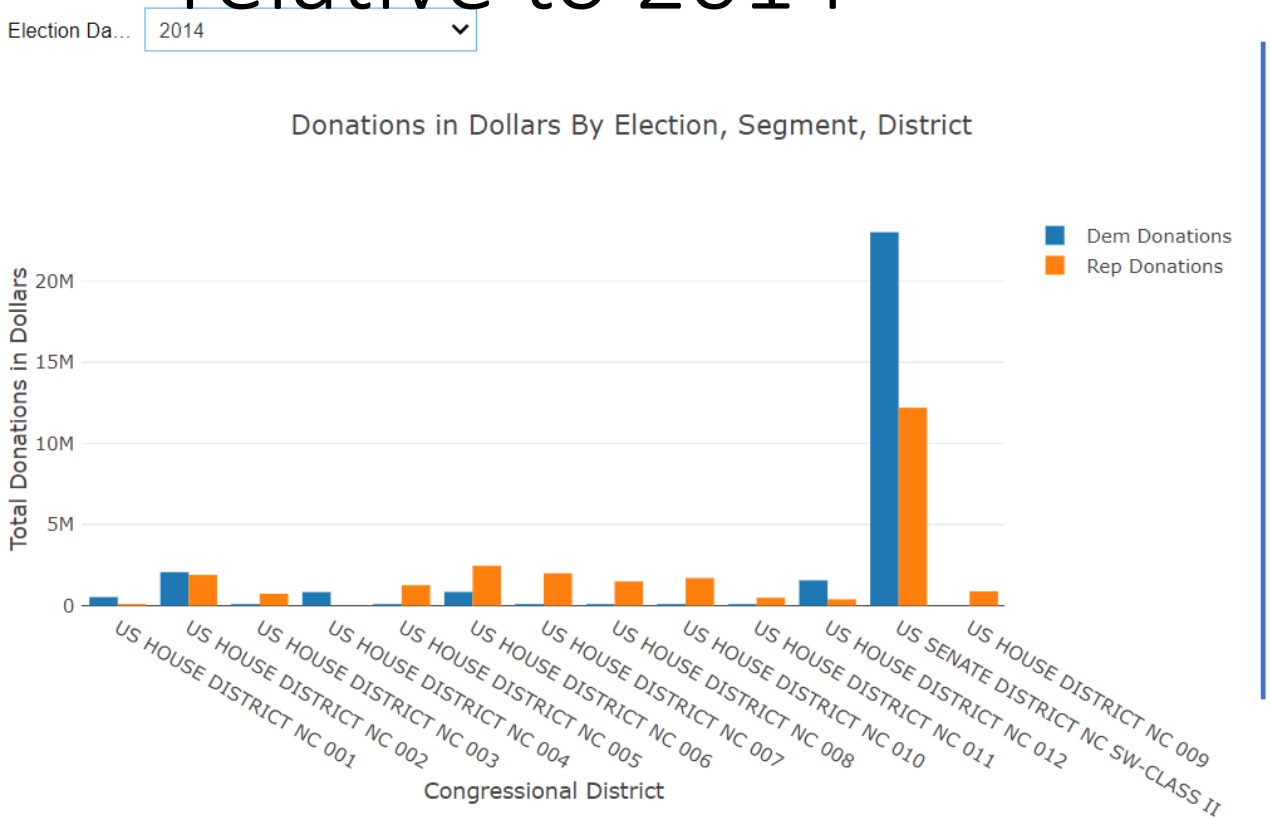
- The sharpest increase between 2014 and 2016 was among Democrats in CD 12 (+50% CoC)
- In 2016, 86.2% of registered Republicans in CD 13 voted (the highest among party/CD combinations. That same year, 74.3% of Democrats in CD 12 voted (the lowest)

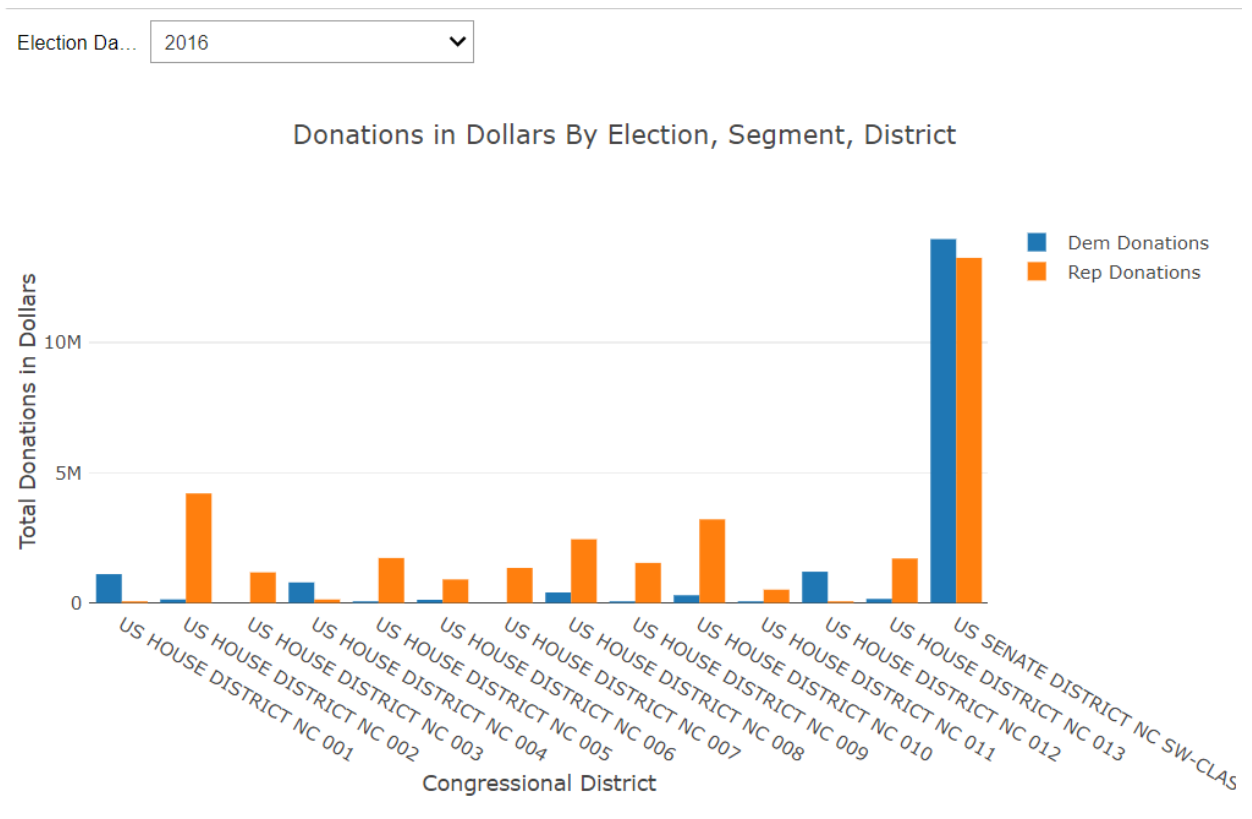
Between 2010-2016, Democratic candidates only received more donations than Republicans(total $) in 2014

- 2014 saw an incumbent Democratic Senator (Hagan) raise $23MM, but lose to challenger Tillis (R) who raised only $10MM
- 2016 included a closely watched Governor's race, which raised $40MM. Cooper (D, challenger) beat McCrory (R, incumbent)



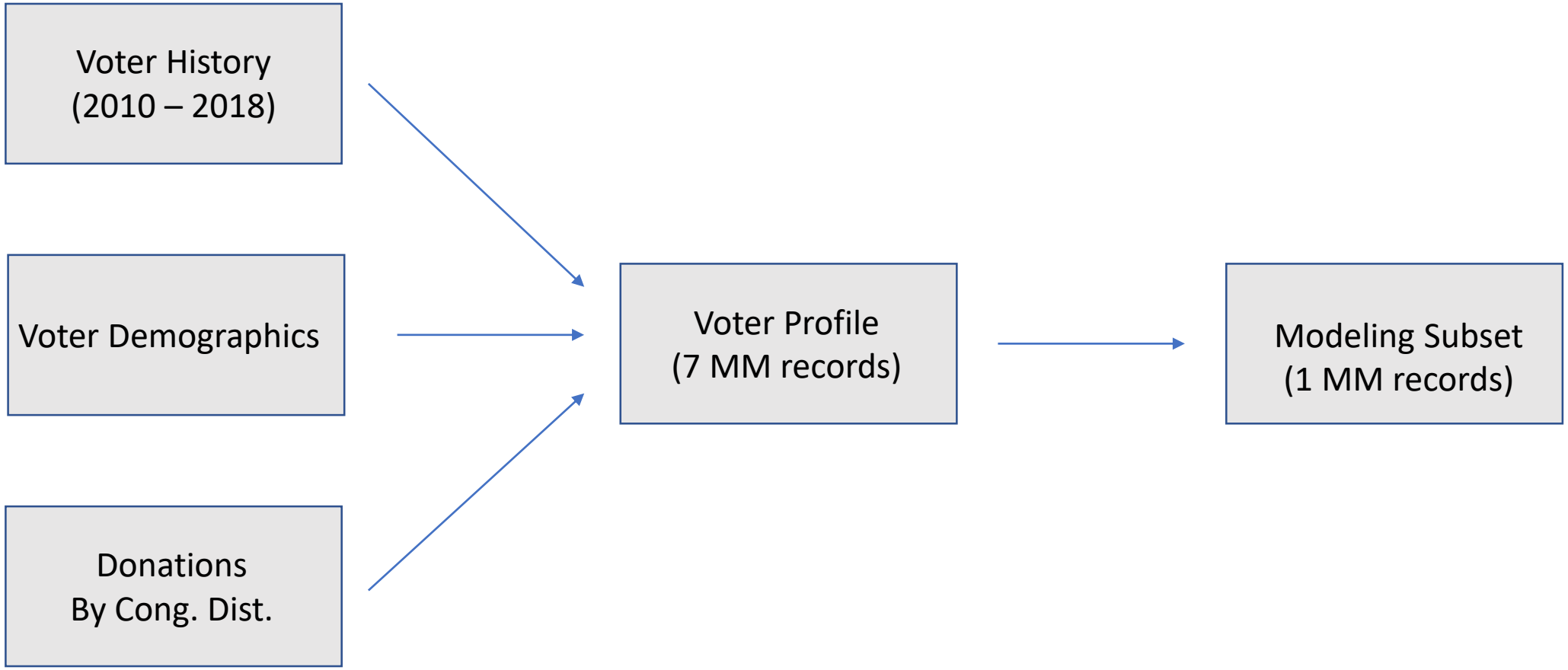
Although voter turnout was higher in 2016, congressional donations were lower that year relative to 2014





- $ 21MM was raised for Congressional races in 2014 vs $23MM in 2016
- On the Senate side $35MM was raised in 2014 and $27MM in 2016

Multiple data sources are needed to create the voter profile



Model Evaluation

- Sensitivity:
  - Among those who voted, how many did we predict correctly?
- Specificity:
  - Among those who did not vote, how many did we predict correctly?
- Accuracy:
  - What percentage of observations did we predict correctly?
- False Negative:
  - We predicted they would not vote, but they did
- False Positive:
  - We predicted they would vote, but they did not
- We want to minimize false negatives
- Optimize model for sensitivity of test set (2016)

Naïve Bayes gave us the best sensitivity score among the test set (2016)

Model	set	pred pos	pred neg	tn	fp	fn	tp	spec	sens
Logistic Regression	train	361,782	638,218	527,024	82,677	111,194	279,105	0.864397	0.715106
	test	389,649	610,351	334,690	31,702	275,661	357,947	0.913475	0.564934
Naïve Bayes	train	459,978	540,022	438,833	170,868	101,189	289,110	0.719751	0.74074
	test	467,991	532,009	268,944	97,448	263,065	370,543	0.734033	0.584814
Random Forest	train	385,375	614,625	596,021	13,680	18,604	371,695	0.977563	0.952334
	test	373,875	626,125	335,566	30,826	290,559	343,049	0.915866	0.541422
AdaBoost	train	382,091	617,909	468,945	140,756	148,964	241,335	0.769139	0.618334
	test	386,341	613,659	276,723	89,669	336,936	296,672	0.755265	0.468226

Donations Not a Valuable Predictor

- Donations and donors by congressional district was not a valuable predictor in the classification models
- The coefficients for these features were close to 0
- Including this data did not improve the performance of the model
- This may be due to a disconnect between voter turnout and donations

Logistic Regression model did not improve substantially with the addition of donation data

Model	set	pred pos	pred neg	tn	fp	fn	tp	spec	sens
Logistic Regression	train	361,782	638,218	527,024	82,677	111,194	279,105	0.864397	0.715106
	test	389,649	610,351	334,690	31,702	275,661	357,947	0.913475	0.564934
LR With Donation	train	362,418	637,582	526,711	82,990	110,871	279,428	0.863884	0.715933
	test	394,765	605,235	333,337	33,055	271,898	361,710	0.909782	0.570873

Naïve Bayes performance declined when the donation data was added

Model	set	pred pos	pred neg	tn	fp	fn	tp	spec	sens
Naïve Bayes	train	459,978	540,022	438,833	170,868	101,189	289,110	0.719751	0.74074
	test	467,991	532,009	268,944	97,448	263,065	370,543	0.734033	0.584814
Naïve with Donation	train	411,035	588,965	390,753	218,948	198,212	192,087	0.640893	0.492153
	test	381,313	618,687	268,972	97,420	349,715	283,893	0.73411	0.448058

Additional data sources may improve model performance

- County level census data
  - ACS 5 year survey
  - Percentage of residents owning homes, income, education levels
- Polling information
  - NC specific
  - High Point University
- Economic Indicators

Thank you!

Questions?