2020 Election Night Model

PPOL 565 - Spring 2021

Madeline Kinnaird

Introduction

The purpose of this project is to replicate, explore, and understand the Washington Post's 2020 Election Night model. Based on the latest reported results, the model's chief objective is to predict the overall turnout at the county-level, thereby bridging the gap between what the results might be showing at any given moment and the likely true underlying result.

Background

Elections and peaceful transfers of power are foundational to a healthy democracy. Therefore, the reporting and tone around election results is critically important. An election night with inconsistent and fluctuating results could undermine American's faith in free and fair elections. Americans are all too familiar with the risks posed from a general election denial sentiment. The 'Stop the Steal' movement was extremely injurious to US election processes and ultimately lead to the violence and domestic terrorism at the US Capitol on January 6th 2021 (McGreevy, 2021). Not only that, it has also been proven that reporting on Election day can impact the liklihood of a person to vote (Jackson, 1983).

Essentially, it is critical that the media's portrayal of election results are tempered, well-communicated, and evenhanded. As such, this project will evaluate the reliability and correctness of the models underlying the Washington Post's 2020 Election Night reporting and work to answer the questions: can we predict outstanding votes on election night and how can we communicate uncertainty in election reporting?

Data

Washington Post's Model Data

In order to closely resemble the model used by the Washington Post on election night, I will use similar county-level data. It is mainly demographic in nature and includes raw counts of census reported categories. The categories are listed and described in Table 1.

While there are other election night models that focus on the precinct level, I do not have access to the data that would enable similar granularity for the entire United States. Additionally, unlike other prominent election forecasting models, the Post's model does not rely on polls. In fact, the model takes no inputs besides previous election results and the demographic makeup of various counties across the United States.

The model we will be inspecting assumes that the county-by-county differences between the 2020 and 2016 turnout are correlated. If we assumed every county's change from 2016 was independent, observing results from any set of counties would not help us predict turnout in counties that have yet to report any votes. Luckily for us, election results are not quite so arbitrary. For example, the change between 2016 and 2020 turnout in Detroit, MI is likely to be fairly similar to the change between 2016 and 2020 turnout in Milwaukee, WI. Our goal then is to build a model that uses final results from one county to help us understand what is happening in other demographically similar counties throughout America.

Data Description

As detailed above, the county demographics data is critical in understanding the correlations of voting outcomes. The data we will use in this project was mostly retrieved from the Census Bureau API using the block level American Community Survey of 2019. The following table summarizes the raw data inputs for the post's model.

Table 1: Election Night Features

Variable Name	mean	std	min	max
Female Count	51723	166564	36	5112188
Male Count	50144	160818	30	4969382
White Population	73778	201567	16	5168443
Black Population	12925	55255	0	1217416
AAPI Population	5568	40938	0	1473221
Hispanic Population	19178	125004	0	4888434
Other Race Population	5155	45040	0	2115548
Median Income	52648	14989	12441	142299
Bachelor's Degree or Higher	20868	67963	8	1962585
Age (<30)	16851	57301	5	1827680
Age (30-45)	19734	68966	12	2155295
Age (45-65)	26302	82313	25	2547857
Age (65+)	15974	45705	11	1335978
Total 2016 Votes	43701	125720	65	3434308
Total 2020 Votes	47755	136034	0	3674850

Because we are interested in inference, namely confidence intervals to directly estimate outstanding votes, our model will use the raw counts listed above. The main thing to note is the enormous variance in magnitudes across all population counts. This introduces a theme that we will see time and time again throughout this project, the rural-urban divide and the extreme differences from county to county.

While our modeling will rely on raw population counts displayed above, we will use normalized values (or percentages) for data exploration; as they are an easier method to spot trends. Table 2 shows our variables as percentages of each county's total population. It is perhaps important to note this distinction, as most of the visualizations in this next section will deal with features in the form of percentages, while the modeling section will utilize raw numeric inputs.

Table 2: Election Night Feature Percentages

Variable Name	mean	std	min	max
Female Count	49	2	27	57
Male Count	50	2	42	72
White Population	82	17	3	100
Black Population	9	14	0	87
AAPI Population	1	2	0	42
Hispanic Population	11	19	0	99
Other Race Population	2	5	0	74
Bachelor's Degree or Higher	19	3	1	40
Age (<30)	14	4	4	52
Age (45-65)	26	2	10	56
Age (65+)	18	4	3	56
Total 2016 Votes	44	8	2	192
Total 2020 Votes	48	9	0	114

Target Variables

As mentioned previously, the Washington Post model functions under the assumption that the county-by-county differences between the 2020 and 2016 turnout are correlated. Since our model forecasts turnout once results start coming in, the quantity we are interested in predicting on election night can be expressed as the difference between the 2020 turnout that we have observed and the 2016 turnout.

$$T^{(20)} - T^{(16)}$$

This project has something that the Washington Post's model did not on election night, hindsight. We now have access to the 2020 Presidential Election returns and are therefore able to observe the actual change in turnout by county. We do just this in Figure 1; inspect a distribution of the percentage of each county's population that voted, comparing 2016 and 2020 on the left and plotting the percentage change on the right.

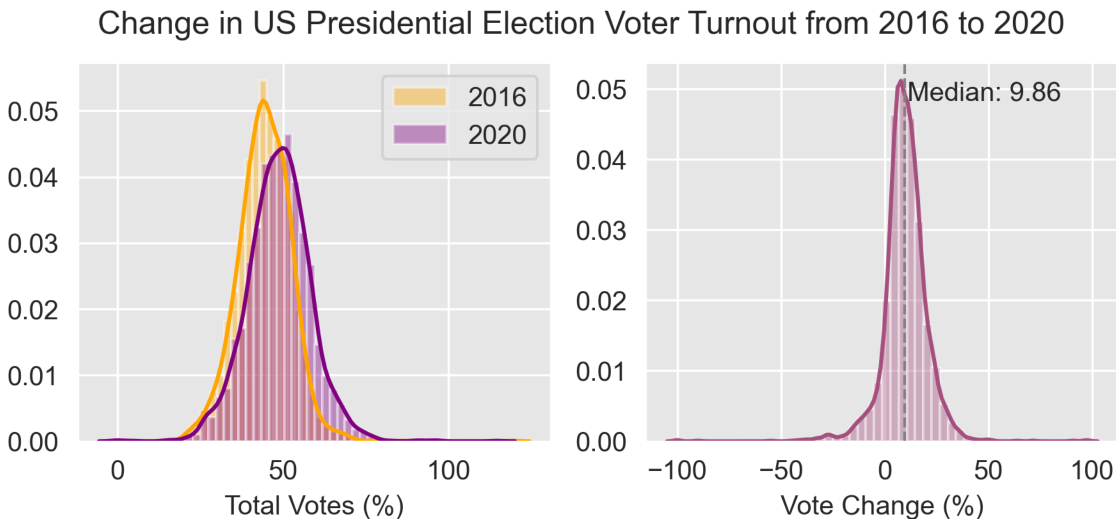


Figure 1: Vote Change between 2016 and 2020

From this figure, we infer that there was a *general* increase in voter turnout from 2016 to 2020 across all counties. The range of percent change by county was varying, but is centered around the median 9.85.

These changes are not uniform however, and can be seen geographically in Figure 2. The yellow counties reflect a high increase of election participation relative to 2016, whereas the purple to gray regions indicate a decrease in participation.

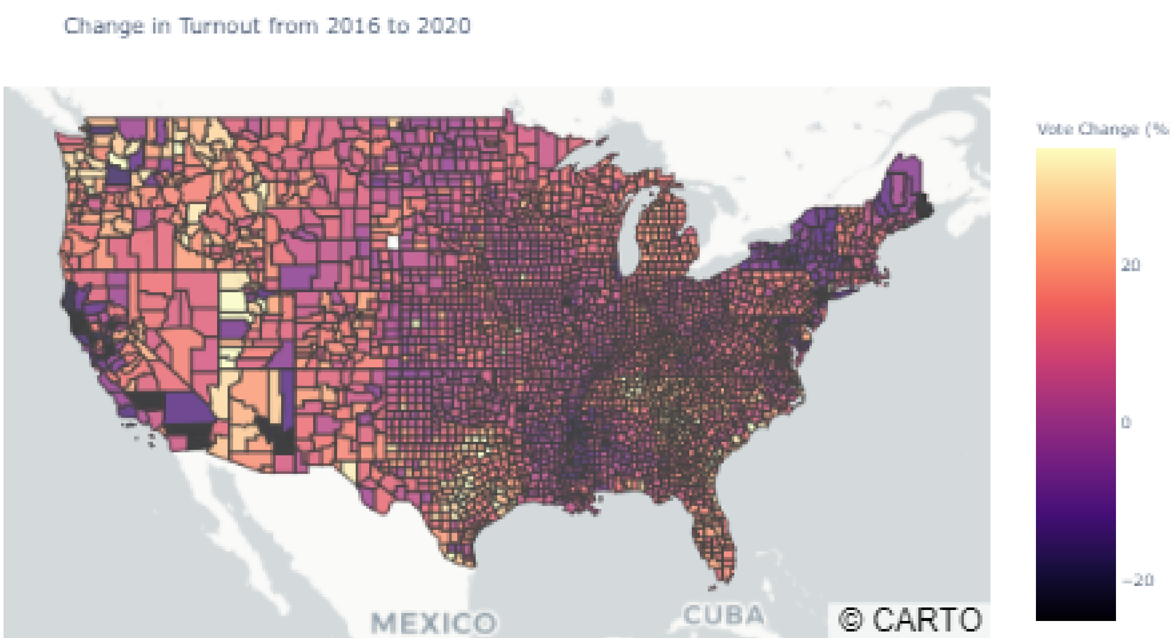


Figure 2: Map of Percentage by County

We note that mainly swing states (such as Arizona, Wisconsin, Pennsylvania, etc) show a increase in turnout while strongly consistently voting states (like California) show a decline.

Additionally, we observe some spatial correlations as many of the trends across counties seem to gradually transition. We note this with a couple of exceptions. There seem to be randomly counties in the middle of the country that show extremely high increases.

This begs the question, what is the relationship between percent increase within a county verses the total population of that same county. In other words, are these instances of high change solely due to a low county population and the notion that individual votes have greater sway to percentage? Therefore; we consider this question in Figure 3.

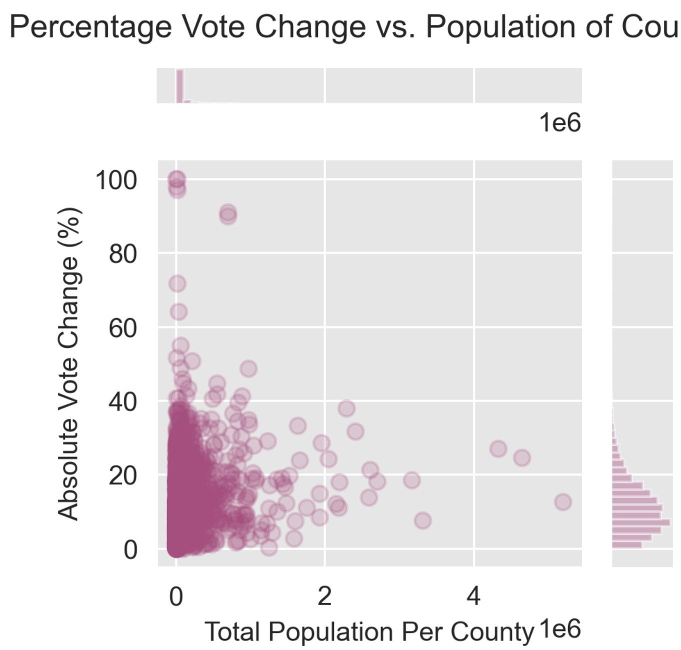


Figure 3: Vote Change vs County Population

From the plot above, we see from the x-axis distribution that an overwhelming number of counties have a rather low total population, this could help explain partially why we are seeing some of the random counties with high changes in voter turnout. Unfortunately, our democratic system was built to favor land over people, so we cannot simply write off these high percentage changes in counties with low populations. We ultimately want a model that will use the final results from one county to help us understand what is happening in other demographically similar counties throughout America. We will next inspect our demographic variables.

Demographic Variables

Through the Census Bureau's API, I obtained county-level demographic data that similarly matched the election night model from the Washington Post. First we'll take a look at the age brackets used in the election night model and how they affect voter turnout in general in Figure 5.

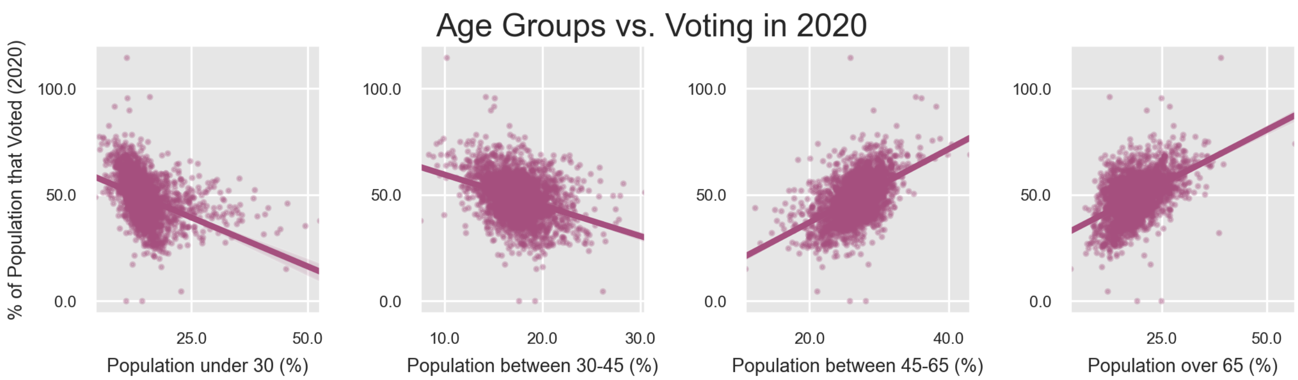


Figure 4: County-Level Age Trends for Voting

We can see that counties with a greater percentage of young adults tended to have a lower percentage of voter turnout in 2020. This matches our intuition as younger demographics stereotypically vote at lower rates than older populations. Next we can look at if these same age brackets drove any substantive change in the voter turnout from 2016 to 2020.

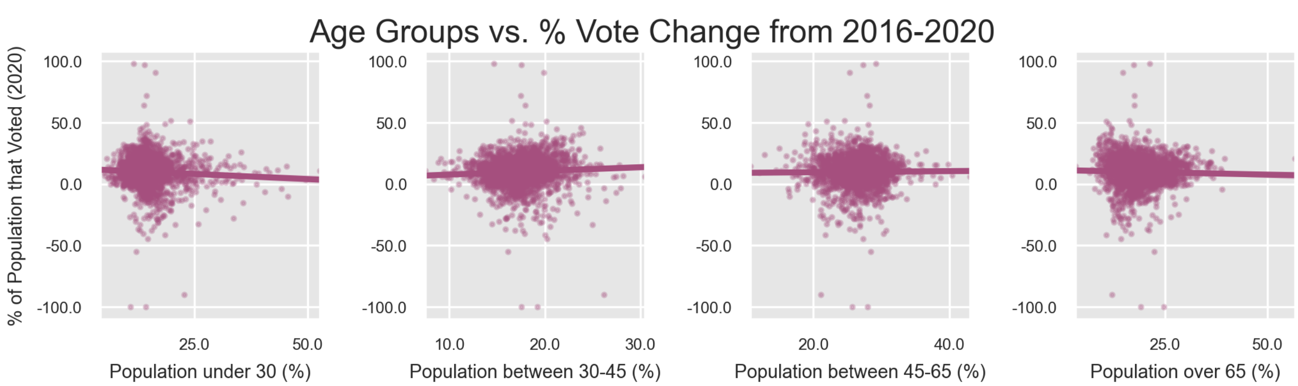


Figure 5: County-Level Age Trends for Vote Change

We note from the above figures, that while age demographics played a notable role in voter turnout *generally* there is not evidence that age demographics played a uniformly significant role in the turnout *change* from 2016 to 2020.

The next census demographic variable we will consider is racial composition of each county. The Census API gives us access to the counts of county residents who identified as any given race. As such, there are multiple ways to consider race categorizations since many people are not singularly one race but multiple. It is unclear how the Election night model categorized racial feature inputs; since they did not post their data and worked with third party data providers. For the sake of simplicity I used the ”checked one box” categorization. Meaning that it is the number of people who checked only one race on the census survey. If we observe the racial percentages of each county, we can learn more about the county distribution across America.

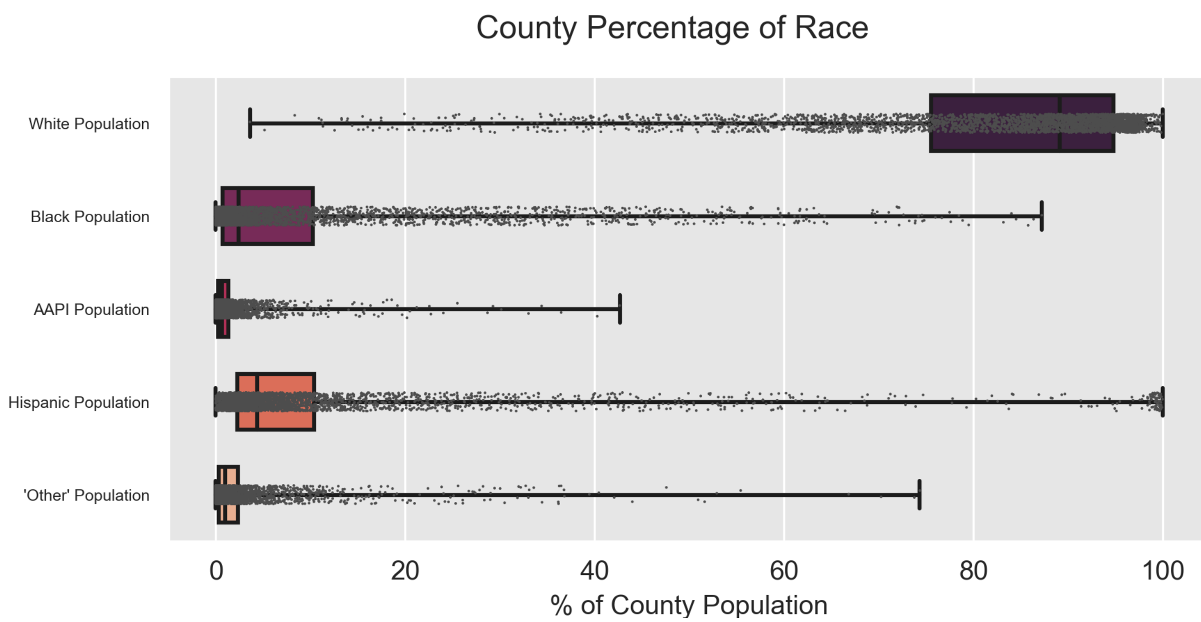


Figure 6: County Racial Composition

We see that there is a rather high number of counties with a majority white population percentage. This could indicate multiple things. First, this could be associated with the “one checked race” Census API issue mentioned above. Additionally, it could be reflective of census participation, as there are multiple studies that indicate the census has a long history of under counting minorities (Sullivan, 2020). Lastly, this could (yet again) be a reflection on the urban-rural divide. Many small rural counties are almost entirely white. Since counties are not divided to have equal populations this white skew could further emphasize the effect of small population counties have within our data.

To investigate the effect of county population size against the percentage of white population within a county, we plot the two in Figure 7.

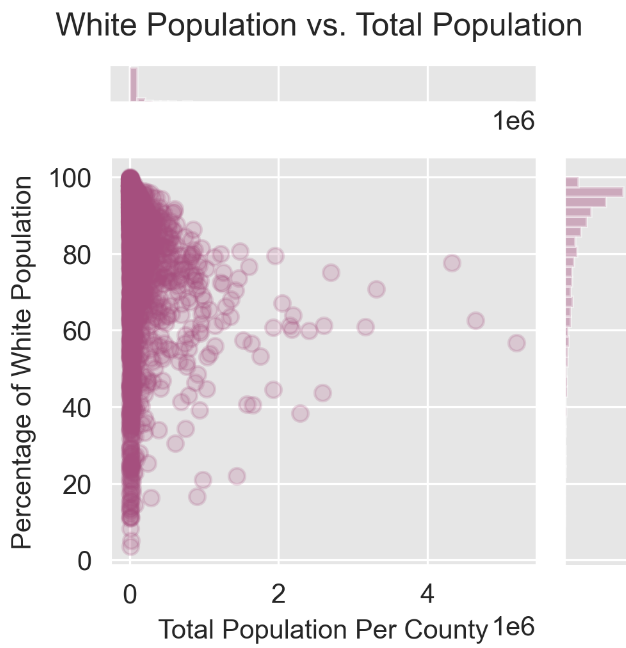


Figure 7: White Percentage vs. County Population Size

From the figure above, we observe that there is indeed a component to the white percentage population skew that has to do with a county's population size. This reiterates the significance of county differences and will be important to keep in mind as we think about modeling and model evaluation.

The last demographic variable we will consider is a county's median income. Figure 8 displays the relationship of voter turnout and distribution of this demographic feature.

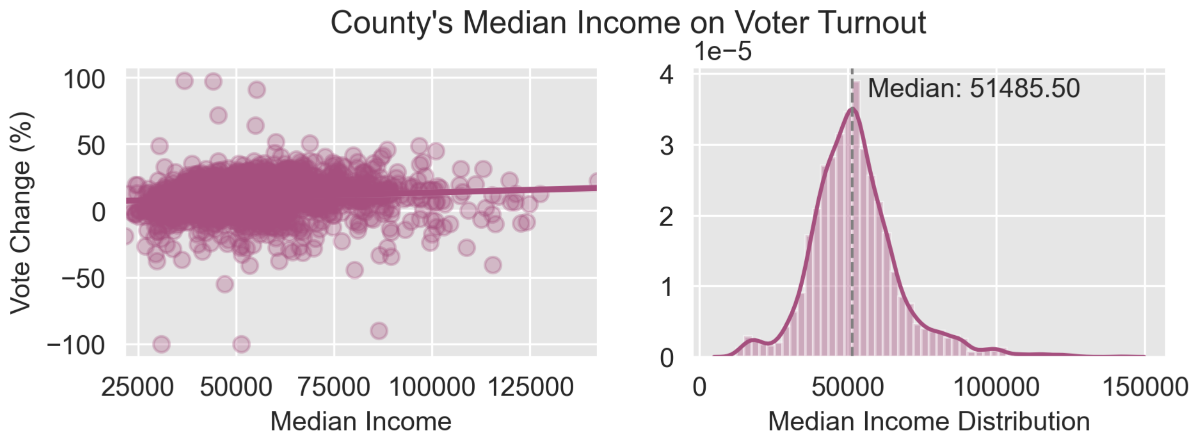


Figure 8: County Median Income

We can see that the median of each county's "median income" (so meta!) is around $51,000. We also observe that median income of a county had a positive effect on 2020 increase in voter turnout. With this, in conjunction with all of the previously listed demographic features, it is our hope to better understand how to proportionately scale the 2016 Presidential returns to accurately predict the returns for 2020.

Methodologies

Analytical Question

A county-level map of vote share says a lot about the politics of America. Vast seas of red rural and exurban counties surround blue islands of metropolitan areas. Critically, it should be apparent that the map is not colored at random. In this project, we aim to determine if the methods utilized by the Washington Post were effective in estimating Election Night voter turnout by leveraging the county-level similarities and demographic composition. Since we are very interested in the confidence interval of our prediction, we are mainly concerned with inference and will therefore consider models that are highly interpretable and output prediction intervals.

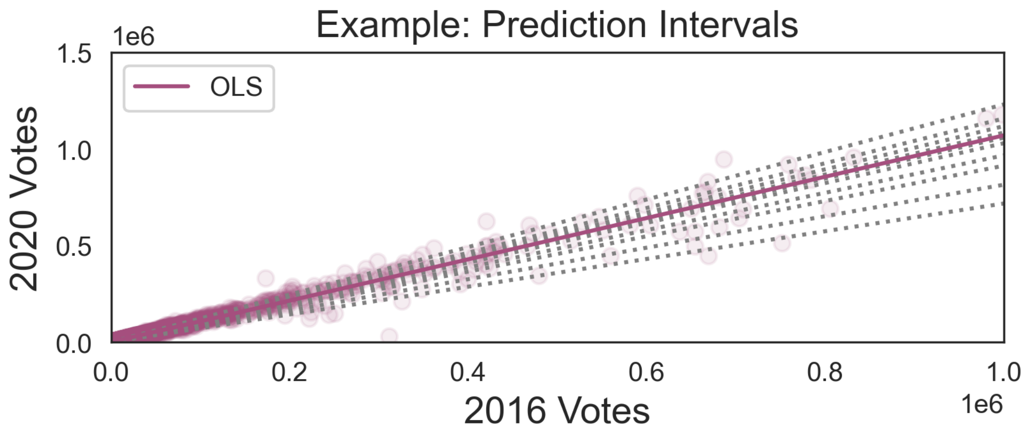


Figure 9: Visual Example of Quantile Regression and Calculating Intervals

I. Gradient Boost Regression

As we ultimately want to predict a numeric value (2020 county votes), a regression model is an appropriate approach to test. The specific method used in this project allows for us to generate prediction intervals by optimizing the quantile loss. For a parallel example, we can look at research conducted by two researchers at Sociologický ústav - Slovenská akadémia vied (Sociológia, 2020). They used this regression to explore the explanatory power of municipal demographics on voter turnout. They were able to use variables like education attainment, economic indicators, health, and many others to predict the percentage of municipal voter outcomes. The found that higher quality of life in a municipality is positively associated with higher turnout. Furthermore, the presence of at least two candidates in down ballot elections increases voter turnout by 10 percent. They were able to use regression to decompose results of voter turnout which is highly applicable to this project.

II. Random Forest Quantile Regression

For a non-parametric option, we will use random forest quantile regression model to estimate the parameters of our model. This will allow us to directly estimate the median as well as lower and upper bounds of our prediction interval without making any distributional assumptions.

An example case that supports the use of quantile regression,focuses on the rational voter hypothesis for Norwegian school language referendums (Kaniovsk, 2013). The experiment was specifically concerned with how the closeness of a race affected voter turnout. They were able to determine that while closeness and size cannot explain the absolute level of voter turnout, they can account for change of the margin. Quantile regression is particularly good at determining intervals making it a very well-suited model for our purposes.

III. Conformal RF Quantile Regression

The conformal part of the model is best thought of as a wrapper that can be put around any black box prediction method to produce valid prediction intervals. The official method we use, "conformalized random forest quantile regression," (try saying that 7 times fast!) wraps a quantile regression model to output a guess for the bounds of a prediction interval.

As CQR is the official model used by the Washington Post, I want to state my intentions and methods more explicitly. I will be using the nonconformist package (donlnz, 2020) to wrap a Scikit-garden random forest quantile regressor model. The goal here is to generate the narrowest interval possible while reaching the desired coverage. The nonconformist package does this by first training an underlying model; in our case quantile regression. Next, we define a nonconformity measure. This is a measure that will indicate how "rare" an observation is compared to the rest of the data. Using this measure, we will calculate the non-conformity scores on a training then calibration data set and determine our desired confidence level. Next, we find the index of the calibration score that corresponds to our desired percentile coverage. This index then helps generate the bounds for our underlying model. Voilá! We have a conformal interval.

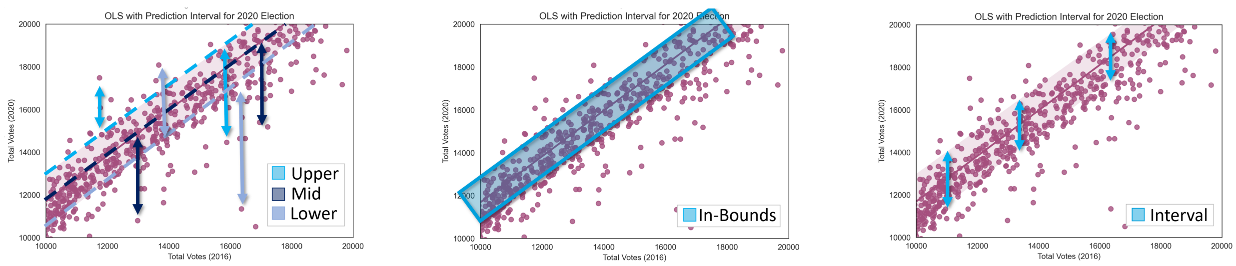
Methology of Evaluations

Next, to evaluate the various models and their intervals, I will be comparing them across three paradigms:

(a) Percent Error: We'll look at the residuals for our predicted values and their upper and lower limits, then normalize those values by dividing by the total number of votes in 2020 for that county (Example A).

(b) In-Bound Percentage: This percentage represents the number of actual 2020 observations that fell within the model's predicted bounds (Example B).

(c) Relative Interval: This metric observes each model's distribution of interval widths; again normalized by dividing each county's interval by the total number of votes in 2020 (Example C).



(a) Percent Error

(b) In-Bound Percentage

(c) Relative Interval

Figure 10: Evaluation Examples

Results

I. Gradient Boosting Regression

Our initial results of the boosted regression model in Figure 11 indicate a couple things. First, it's important to orient ourselves in these metrics. The lower bound should consistently have negative errors (and vice versa) as the residuals are relative to the true 2020 voter turnout; essentially we want our lower bound to have negative errors and our upper bound to have positive errors, with the mid error as close to zero as possible.

For this model, the upper boundary was the most inconsistent with errors 300% the actual number of county votes. The lower boundary has a fairly narrow range of error percentages.

Additionally, when we look at the prediction interval generated by the model we can see that most of the upper boundary errors occurred in counties with smaller voter turnout (the bottom left corner) and most of the lower boundary intervals occurred in the counties with larger voter turnout (the top right corner). Polarized-looking errors generally make sense as the x-axis scale is logged and these outlying sections can mostly be considered outliers; although the clear and opposite directions of the outlier's residuals is interesting to note.

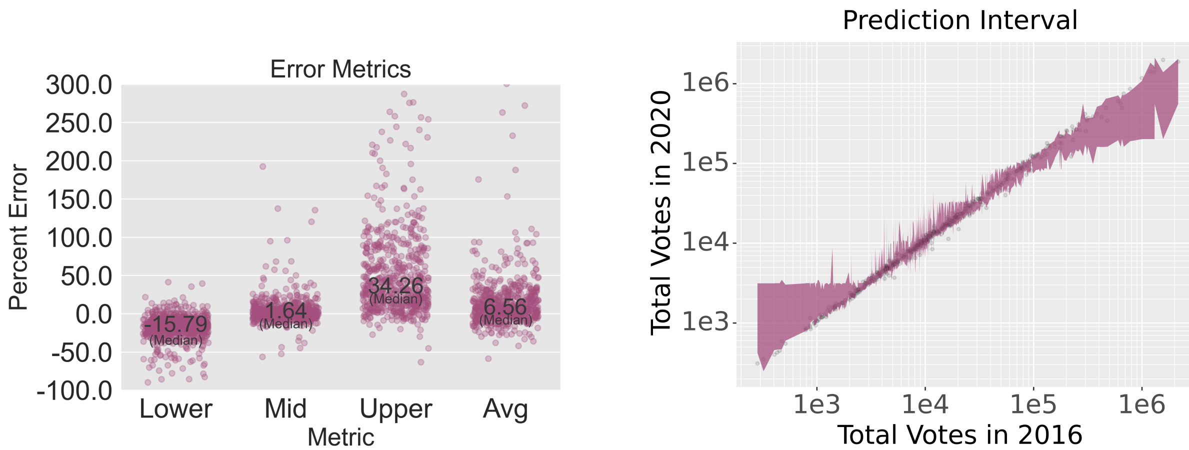


Figure 11: GB Regression Results

II. Random Forest Quantile Regression

The results of the Random Forest Quantile Regression model has a more consistent range of errors, with the general percentages spanning only about 22%. Looking at the RFQR's prediction intervals, we see that once again the the outlying regions have wider intervals. This time, the cases of outliers seems to be associated with lower bound errors. This is helpful in understanding how each model handles extreme cases.

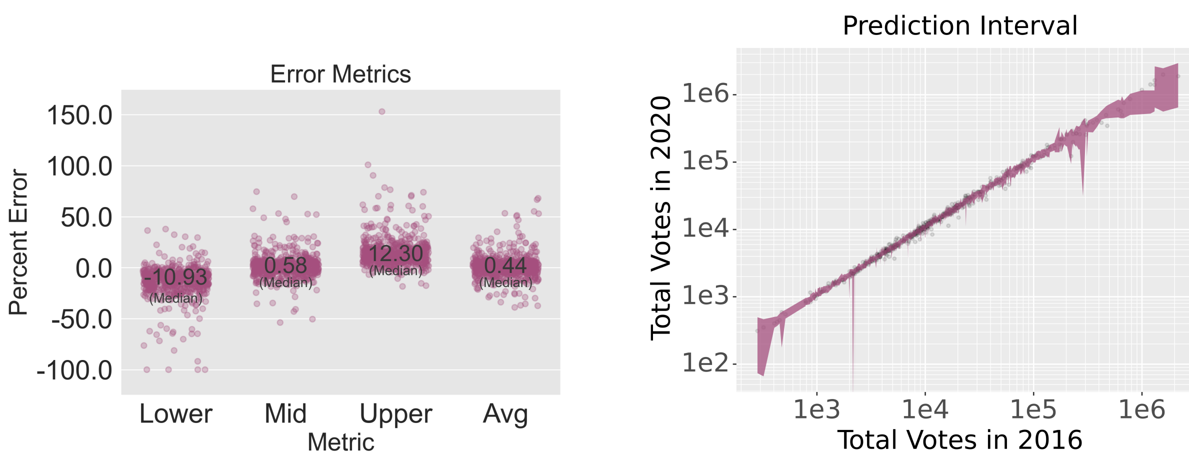


Figure 12: RFQR Results

III. Conformal Quantile Regression

And last but not least, we observe the performance of the conformal quantile regression model. Here we see that there is a rather wide range on both the upper and lower boundaries. This aligns with our knowledge of this model - remembering that conformal prediction is a wrapper that generates intervals with a high probabilities of containing the true target variable. This could account for the wide ranges of residuals especially on the upper and lower bounds, since the wider the interval the more likely it is to include the actual value. Additionally, CQR predicts successively which could be another aspect of these performance ranges; the order of accumulation could mean the high-error counties were earlier test-points in the model. Lastly, looking at the prediction intervals, there does not seem to be systematic skews based on these plotted axis - which differentiates it from the previous two models.

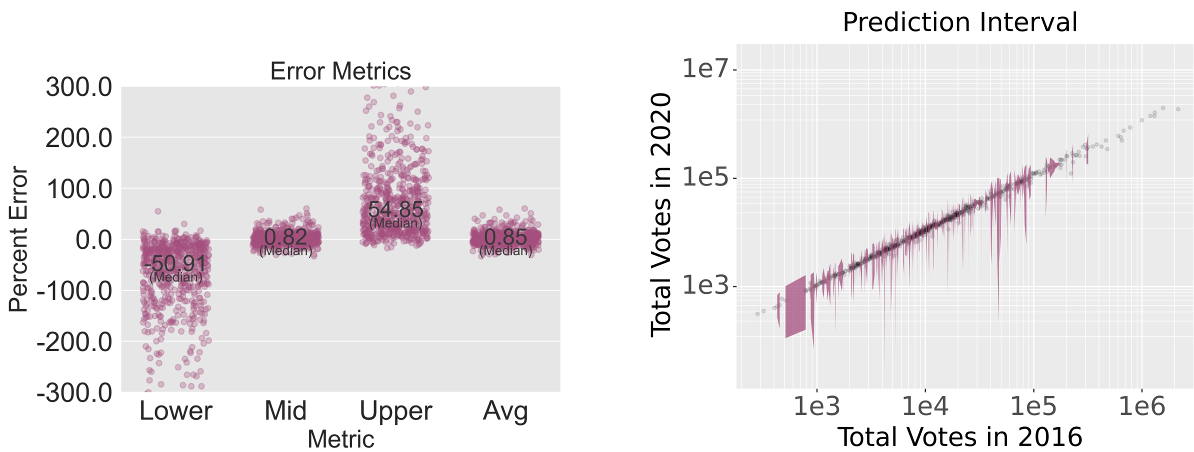


Figure 13: CQR Results

So now that we've had a cursory glance at each model's individual performance, we will compare them across the three metrics listed in the "Methodology of Evaluation" section.

Percent Error For Each Model

Here we compare percent error side by side. The GBF's errors disqualify it from "winning" this round, as it consistently under-performs. Interestingly, CQG has the most precise mid-point prediction while also the widest ranges on it's upper and lower bounds. Reasons for these high CQR bound rates have been previously speculated, but here we see just how much greater these errors are compared to the other models.

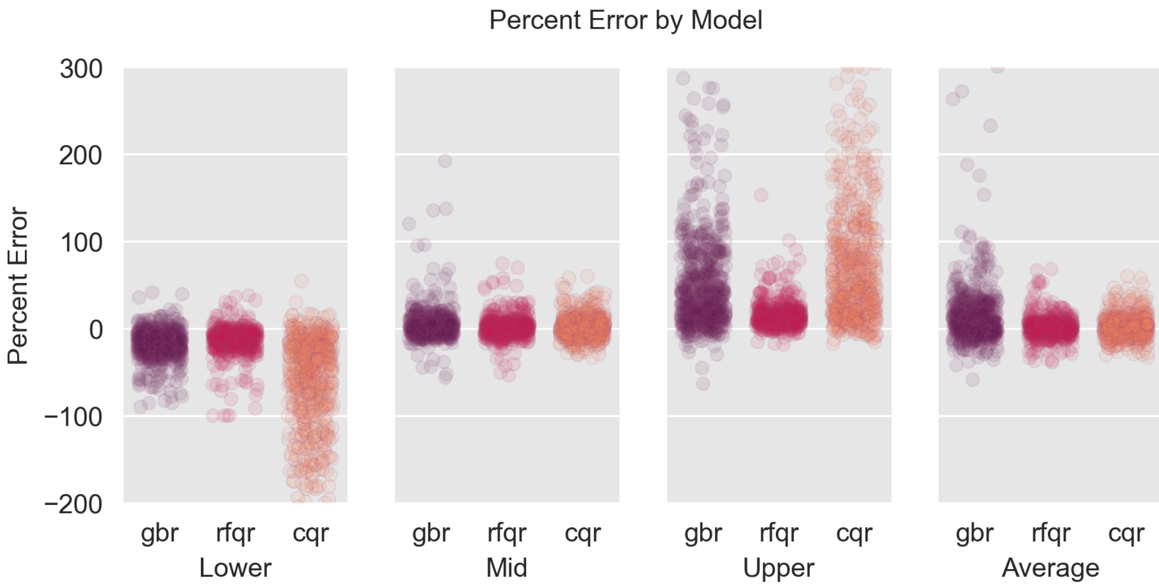
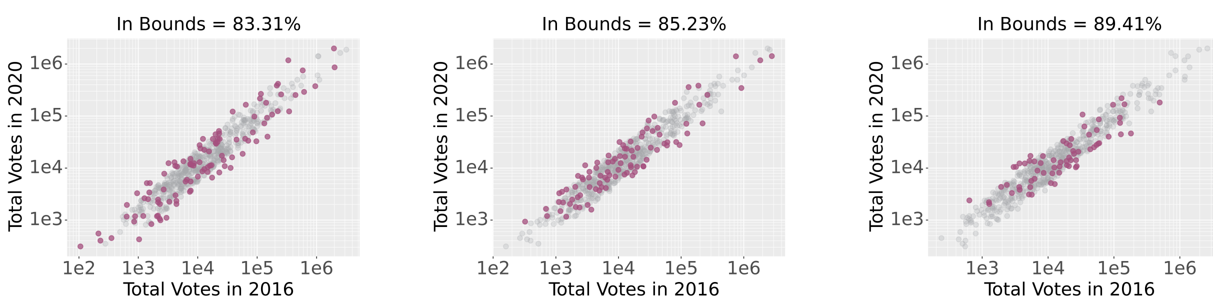


Figure 14: Percentage Error Results

In-Bound Percentage

Below we see each model's performance on a binary spectrum; whether the model correctly predicted a county within it's interval bounds. The "out-of-bounds" counties are highlighted below in magenta while the overall percent of accurately predicted counties are at the top of each graph.



(a) GBR (b) RFQR (c) CQR

Figure 15: In-Bound Percentage Results

From a very high level, we would say that the CQR model performed with the highest level of accuracy based on the 89.41% "in-bounds" rate. But as we know from our analysis of error percentages, a pure accuracy rate does not necessarily mean an obvious winner for best suited model, it might just mean wider prediction intervals. We will look at these interval ranges in the following section.

But before we look at our last evaluation metric, one other thing to point out regarding in-bound percentages is the out-of-bounds distributions of each model. We'll consider the distribution from the perspective of population as the x-axis "Total Votes in 2016" is somewhat a proxy for a county's magnitude (and alreay shows us some patterns among the models). The GBR's out-of-bounds counties seem to span across all types of counties while the RFQR and the CQR's out-of-bounds counties seem to steer clear of the extreme county outlier regions, clustering in the middle. Another outlier observation to consider in our evaluation.

Relative Interval

Lastly, we will look at relative intervals. Again, this is essentially considering how wide each generated model's prediction intervals were. To normalize these numbers and account for the vast magnitude differences of the urban-rural divide, each county's prediction interval was divided by it's total number of votes in 2020. The outcome is a ratio that we'll call a "relative prediction interval". The normalized ratio's number isn't particularly significant by itself, but it allows us to more easily compare different kinds of counties.

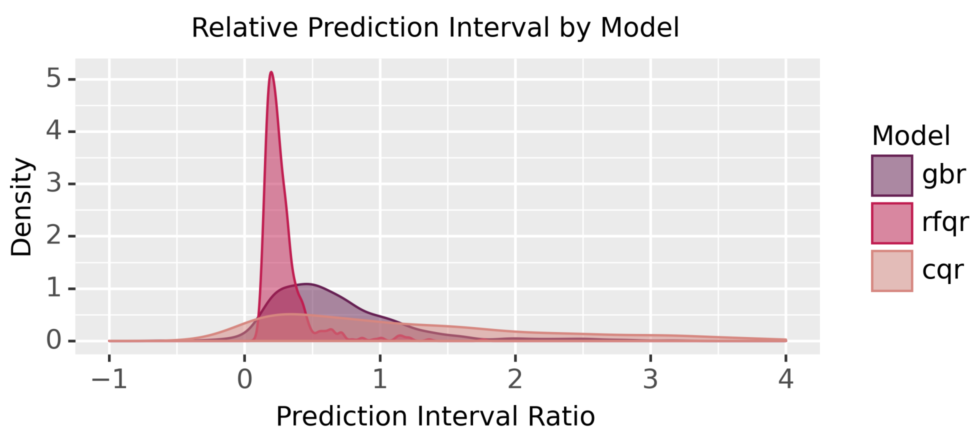


Figure 16: 2 OLS

Here we learn a couple of interesting things: The first is that the Random Forest Quantile Regression model does not dip below zero. That is great! The intervals were obtained by subtracting the predicted upper bound by the predicted lower bound, so a negative number implies a wonky prediction interval. The RFQR model staying above zero bolsters the model's credibility as being fairly robust. As speculated before, the conformal model's interval is the wide range spanning from just below zero to a relative prediction interval ratio of four. This confirms our suspicions of a wider prediction interval.

Conclusion

Implications

When considering what kind of model is best suited for election night, we want a model that is accurate, that affords us an informative interval, communicates uncertainty reasonably, is robust, and not highly susceptible to outliers.

Based on the results we've observed, the best choice is between the Random Forest Quantile Regression Model and the Conformalized Quantile Regression Model. CQR has 4% greater in-bounds rate which could be critical on election night, especially as nail-biting swing states incrementally report their results. On the other hand, the unwrapped RFQR model's underpinnings provide a more robust approach to intervals with only non-negative intervals generated and a much narrower interval for only a 4% loss in "in-bounds" accuracy.

But overall, as mentioned before, we want a robust model, one that does not fluctuate highly on election night and avoids giving nonsensical predictions. Based on the overall performance of the models in this specific instance, I would be inclined to use the Random Forest Quantile Regression model. But with more time to properly calibrate, tinker, and add the proper backstops in place, the conformal prediction wrapper would likely be a nice addition to boost accuracy.

Further Considerations

On the whole, all three models tended to slightly overestimate voter turnout across all counties, but of course the errors are not uniform. Given this current analysis, my next steps and line of inquiry would be to analyze the types of errors occurring. A potential next step for future evaluation of model performance, would be to evaluate the different types of errors made. Are there certain types of counties or certain attributes of a county that causes it to have a particular error? Maybe the models tend to overestimate counties in the south or counties with a large senior citizen population. If we were worried about those kinds of errors we could run a clustering algorithm to see if there are systematic errors the models are making and further analyze that state of elections in America for 2020 and beyond.

Bibliography

Bureau, US Census. “American Community Survey 5-Year Data (2009-2019).” The United States Census Bureau, https://www.census.gov/data/developers/data-sets/acs-5year.html.

Jackson, John E. “Election Night Reporting and Voter Turnout.” American Journal of Political Science, vol. 27, no. 4, 1983, pp. 615–35. JSTOR, doi:10.2307/2110886.

Kaniovski, Serguei. “A Test of the Marginalist Defense of the Rational Voter Hypothesis Using Quantile Regression.” In Power, Voting, and Voting Power: 30 Years After, edited by Manfred J Holler and Hannu Nurmi, 87–102. Berlin, Heidelberg: Springer, 2013. doi.org:10.1007/978-3-642-35929.

Linusson, Henrik. Donlnz/Nonconformist. Python, 2021. https://github.com/donlnz/nonconformist.

McGreevy, Nora. “Was the Capitol Attack Part of a New Wave of Terrorism?” JSTOR Daily, 16 Feb. 2021, https://daily.jstor.org/was-the-capitol-attack-part-of-a-new-wave-of-terrorism/.

MIT Election Data And Science Lab. County Presidential Election Returns 2000-2016. Harvard Dataverse, 2018. DOI.org (Datacite), doi:10.7910/DVN/VOQCHQ.

“Socioeconomic or Political Variables? The Determinants of Voter Turnout in Czech Municipalities.” Sociológia - Slovak Sociological Review 52, no. 3 (2020): 222–44.

Sullivan, Teresa A. “Who’s Missing? Undercounting and Underreporting.” In Census 2020: Understanding the Issues, edited by Teresa A. Sullivan, 33–47. Cham: Springer International Publishing, 2020. https://doi.org/10.1007/978-3-030-40578-.

Washingtonpost/2020-Election-Night-Model. 2020. The Washington Post, 2021. GitHub, https://github.com/washingtonpost/2020-election-night-model.