Predicting the Outcomes of the 2020 Election

Victor Avram, Blake Bullwinkel, Teresa Datta, Kristen Grabarz

AC209a, December 2020

1 Motivation & Context

The US election is an important event during which voters have the ability to exercise their voices in determining local and national leadership. Even beyond our borders, the results of the American presidential election have a significant impact on political developments across the globe. For a combination of reasons including underestimated uncertainties and unreliable polling data, the vast majority of election experts failed to predict the outcome of the 2016 election.[8].

The 2020 election came in the midst of unprecedented social and economic conditions in the US. Given the devastating effects of the coronavirus pandemic and the social unrest that followed the murder of George Floyd, the voting outcomes of Americans this year are perhaps more unpredictable than ever. Despite these challenges, in this project, we used data preceding November 3rd, 2020 to build a model predicting the outcomes of 2020 presidential and congressional (House of Representatives) elections. Using historical election data starting from 1976, polling data, and "fundamentals" data (economic growth, presidential approval ratings, etc), we trained and evaluated multiple machine learning models and compared our predictions to the actual election outcomes.

We were also interested in exploring the possible effect of the COVID-19 pandemic on this year's results. We accomplished this by analyzing the relationship between factors such as unemployment rate (a strong proxy for economic impact of COVID-19) and per-capita COVID-19 case rates.

2 Description of Data

- U.S. Census Bureau 116th Congressional Districts and State ShapeFiles: Cartographic boundary files of the 116th U.S. House of Representatives congressional districts and the 50 states.
- US Census Bureau American Community Survey: Contains information on demographic makeup of populations, including race, age, and education levels.
- MIT Election Lab House of Representatives Results 1976- 2018: Candidate general election results for all U.S. House races from 76-18.
- MIT Election Lab Presidential election results 1976 - 2016: Tally of votes cast per party for each state
- FiveThirtyEight presidential election polling averages 1972 - 2016
- Cook Political 2020 Presidential and House of Representatives Election results
- Unemployment Data: Bureau of Labor Statistics, includes monthly unemployment rate by state
- COVID-19 Data: CDC COVID Data Tracker

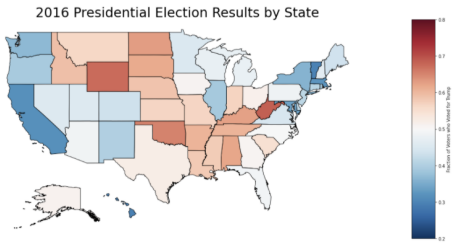

Figure 1: 2016 Results by State

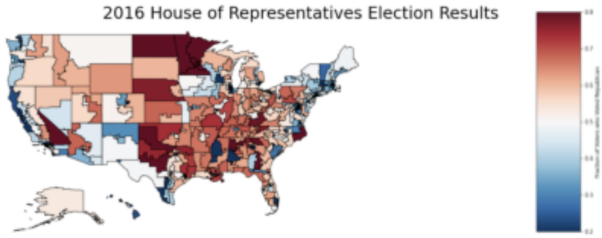

Figure 2: 2016 House Results

3 Exploratory Data Analysis

Figure 1 captures the results of the 2016 election, with the shade of each state indicating what percentage votes were captured by Trump. How firmly "red" or "blue" a state was in 2016 sets the political backdrop for the 2020 race in terms of which states are potential "swings" (the ones with shades closer to white) and which states are unlikely to deviate (the more brightly colored ones).

Similarly, Figure 2 displays the results of the 2016 House of Representatives elections, with the shade of each district indicating the percentage of votes captured by the Republican candidate. Similar to the POTUS map, this map helps indicate which seats are more competitive than others. In addition, this map shows how varying the land area is for each house seat. As we know, the U.S. population is unevenly concentrated in a few major cities.

Figure 3 represents the vote fold change per state across the presidential elections spanning 1976 to 2016. Vote fold change is defined as the percentage of the vote attributed to the Democrat candidate divided by the percentage of the vote attributed to the Republican candidate. A fold change greater than 1 indicates that the Democratic candidate has won the given state and a fold change of less than 1 indicates that the Republican candidate has won the given state. The dashed blue line indicates a fold change of 1. We can see from the given plot which states have historically voted in a certain direction and the variability in voter behavior across time. For example, we can see that Arkansas has always voted Republican since the 1976 presidential election and New York has almost always voted Democrat since the 1976 presidential election.

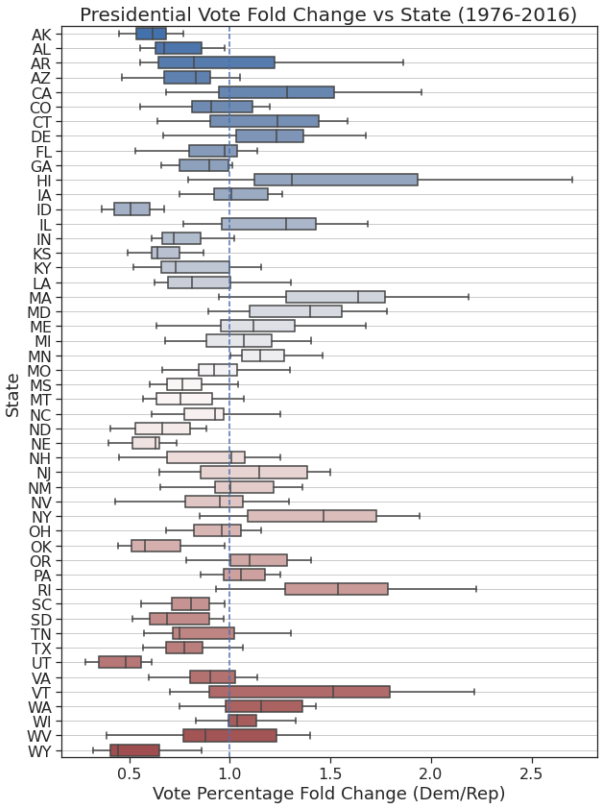

Figure 3: Presidential Vote Fold Changes

We are also aware of the fact that some states and regions of the US are "consistently red" or "consistently blue". For example, the Democratic candidate for president has won Massachusetts in all but two presidential elections since 1972. As Figure 3 suggests, previous election outcomes might be a promising predictor for our model, and we graphically explored the relationship between each state's previous and current election results. Note that the derived variable "fold change" refers to the percentage of the Democratic vote divided by the percentage of the Republican vote. This attempts to account for the fact that third-party candidates are sometimes in the race, so we cannot set a binary cutoff at 50% of votes to determine an election winner.

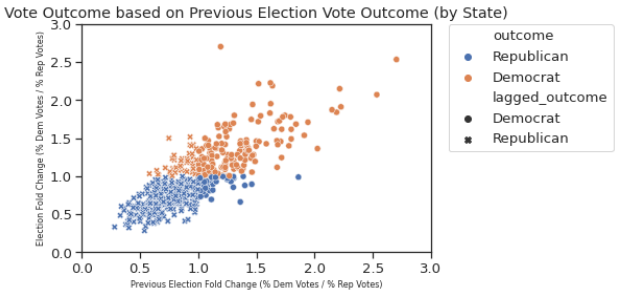

Figure 4: Presidential Lagged Outcomes

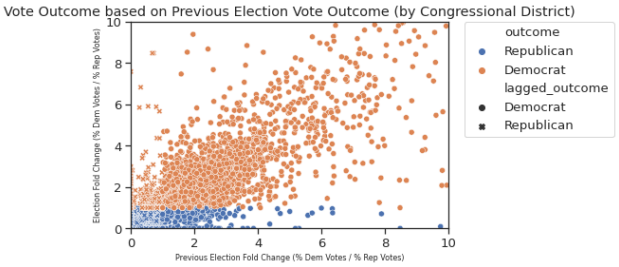

Figure 5: Congressional Lagged Outcomes

As expected, there is a positive relationship between these two outcomes. For good measure, we also lagged the second-most-recent election outcome for our modeling. We explored a similar phenomenon for Congressional districts and found a positive correlation as well, though the trend is a bit looser than for the presidential outcomes (however, the sample size is also smaller: in the case of the presidential elections, we have one observation per state since 1976, whereas in the Congressional version we have one observation per district since 1976). These graphical insights are summarised in Figures 4 and 5 and corroborate our hypothesis that historical election outcomes clustered by state or district may be worth including in a baseline model.

During our background research, we also discovered that political scientists often speak of "fundamentals": structural factors that influence voter decisions such as the state of the economy or whether an incumbent is running. Fundamentals-based forecasting tends to be quite stable, and a classic example is known as the "Time-for-change" model, which was created by Alan Abramowitz of Emory University. This model predicts the popular vote winner using only the president's approval rating, GDP growth, and whether or not a first-term president is running for re-election, and was correct in every election from 1988 until 2016.

Inspired by this simple model's success, we decided to explore fundamentals and see how they might fit into our model. Figure 6 shows the relationship between the percent of the popular vote for a Democratic candidate and the second quarter GDP growth from election year, as well as whether or not there was an incumbent (and a Democratic incumbent, specifically) in the race. Figure 7 shows the same effect in relation to the approval rating of the sitting president in June. Interestingly, the data shows some evidence of a weak positive relationship between the percent of Democratic votes and both second-quarter GDP growth and June approval rating when there is a Democratic incumbent, and the opposite effect when there is a Republican incumbent. This suggests that these predictors may also be useful to include in our model.

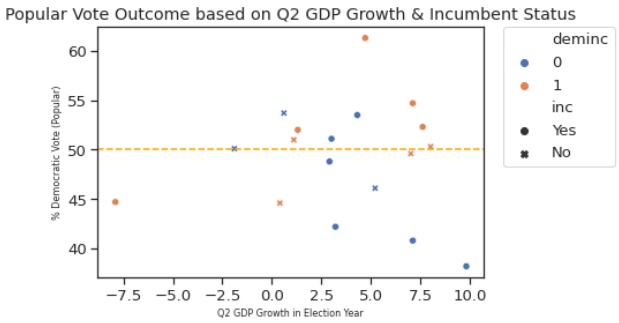

Figure 6: Fundamentals: Approval Rating

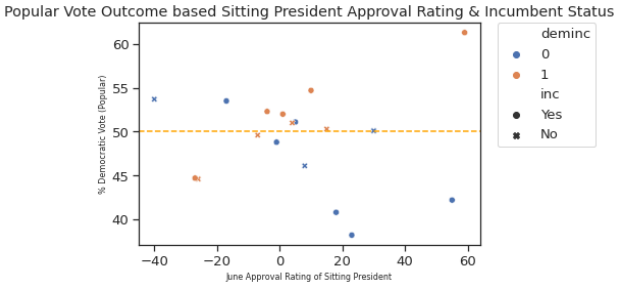

Figure 7: Fundamentals: Q2 GDP Growth

Further, we also considered demographic attributes of the voting population. Using data from the 2016 election and state-wide population metrics from the US Census, we plotted various demographic attributes including age, population, and race, against the percentage of the vote that is democratic. Figures 8 and 9 display the two most interesting visualizations yielded by this exploration. Based on data from the 2016 election, we found that there are positive correlations between population density and education attainment

and the percent of a state's 2016 vote that was Democratic. Intuitively, this makes sense because cities and more urban areas typically lean Democratic, and tend to have dense populations and large proportions of highly educated residents. Taken together, there is a clear threshold between these two variables, above which point most states voted Democratic in the 2016 election. This suggests that demographic information such as education or population density are also valuable predictors to incorporate into our model.

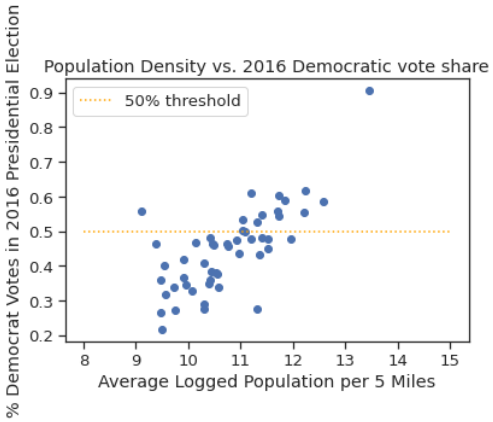

Figure 8: Demographics: Population Density

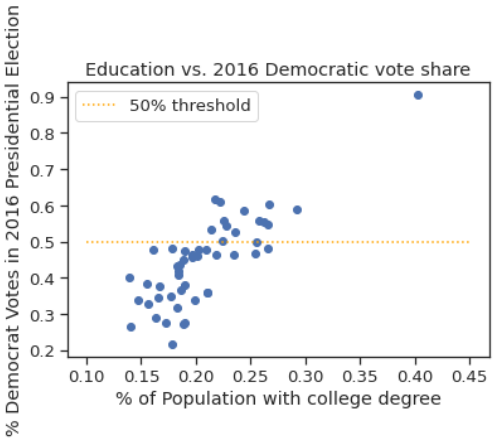

Figure 9: Demographics: Education Level

Another useful source of information is polling data. Particularly for models that attempt to predict outcomes in the weeks and months leading up to an election, polling data can provide a useful pulse on voter preferences. Given the nature of our historical model, we decided to incorporate polling averages for each state, averaged over the months (depending on the availability of data) leading up to each election from 1972-2020. For example, Figure 10 shows the percentage of respondents favoring Biden and Trump over the weeks leading up to the 2020 election in Georgia and New York.

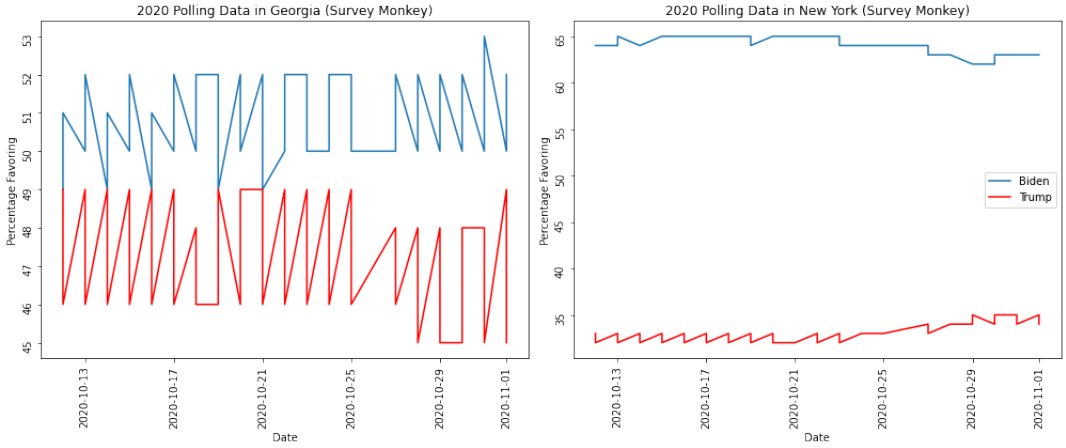

Figure 10: 2020 Polling Data in Georgia and New York

During our EDA, we inevitably encountered missing data. It has been shown that election outcomes for certain states can be accurately predicted from the outcomes of other states that are closely related across elections. Figure 11 shows a state by state correlation matrix, where each element represents the correlation between the vote fold changes (defined as the ratio of % Democratic votes to % Republican votes) of the corresponding pair of states. Figure 12 shows a similar state by state matrix, where each element corresponds to the coefficient of determination (R^2) between the vote fold changes of the corresponding states. This matrices allowed us to perform better-informed imputation of missing election results.

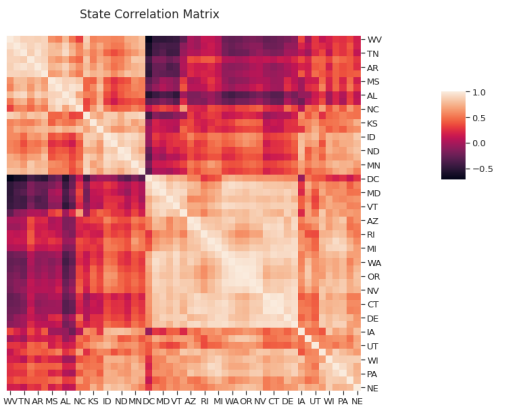

Figure 11: State Correlation Matrix

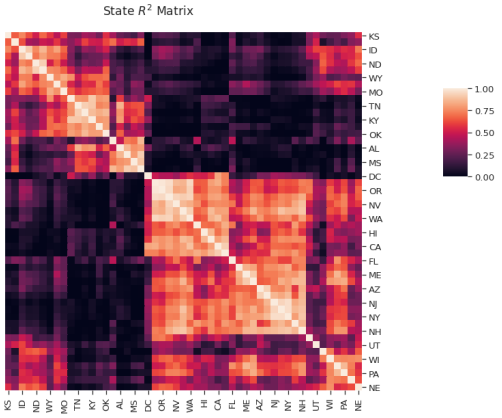

Figure 12: State R^2 Matrix

4 Modeling Approach

4.1 Baseline Models

For a baseline, we developed a model based on variables that seemed promising after our EDA and research into political science research methods. The predictors we used are a combination of historical election outcomes, polling data, and key inputs of the "fundamentals" model mentioned above.

Baseline models were created using data related to presidential and House of Representatives elections spanning the years 1976 - 2016 (11 presidential elections and 22 House elections). For the presidential election, each observation in our dataset was specified by a unique state and year pair (n = 561), and for the Congressional election, each was comprised of a unique district and year pair (n = 9735). These models were built to predict the outcome for a given state or district for the given election year. For predicting the presidential election outcome, the overall outcome for the given year was subsequently derived from the predictions across all states. The predictors used were either state and year specific (e.g. the total umber of votes, the lagged fold change) or year specific (e.g. the end of year GDP, the second quarter GDP). A description of the predictors used in these baseline models is given below.

- **q2gdp**: The second quarter GDP the election year
- **inc**: A boolean indicating whether an incumbent president is in the presidential race
- **juneapp**: Net approval rating of the sitting president based on the Gallup poll conducted at the end of June of election year
- **lagged_FC**: The fold change designated as $\frac{percentageDemocratvote}{percentageRepublicanvote}$ for the previous election cycle
- **twice_lagged_FC**: The fold change designated as $\frac{percentageDemocratvote}{percentageRepublicanvote}$ for the election cycle that occurred 2 cycles prior

The response variable was a boolean indicating whether the winning candidate was a Democrat (1) or a Republican (0). The assumption used for the baseline models and subsequent finalized models is that either a Democratic or Republican candidate will win the election. We do not take into account the low probability case of a candidate from another party winning the given election.
An unregularized logistic regression model was built using the specified predictors. The training set consisted of 80% of the data. The classification accuracies for the training set and test set were 0.83 and 0.82 for the presidential model, respectively. A k-nearest neighbors (k-NN) model was also built to predict the fold change. The same predictors were used as with the logistic regression model. The optimal value for k was determined through an iterative process of testing values of k between 1 and 20. 3-fold cross-validation was used to assess the mean squared error (MSE) on the validation sets, and the model that had the best

performance on the validation sets used $k = 1$. The validation set MSE for this model was 0.790, and the R^2 value and MSE on the test set were 0.967 and 0.052, respectively.

For the Congressional model, a similar methodology yielded a train classification accuracy of 84.3% and a test classification accuracy of 84.5%. A KNN model predicting raw fold change for the congressional scenario used cross-validation to select an optimal value of k. This method yielded an initial R^2 of 0.6 for the House of Representatives with an optimal $k = 17$.

4.2 Model Refinement

Though our baseline models performed fairly well, we sought to improve them further by incorporating additional predictors and more robust modeling techniques. To supplement the variables noted above, we added an additional predictor to reflect polling trends: poll fold-change, which was the average fold-change across all polls administered for a given state or district.

Further, we sought to mitigate model overfitting by utilizing regularization techniques, including both LASSO and Ridge-regularized methods, and we used 3-fold cross-validation tune the regularization parameters. After evaluating the performance of the tuned models on the out-of-sample test set (preceding the 2020 election), the Ridge-regularized model had the greatest performance, with a test accuracy of 0.832 and a c-value (determined via cross-validation) of 10. Below, we summarize performance of this final model on predicting 2020 election outcomes.

5 Model Conclusions & Limitations

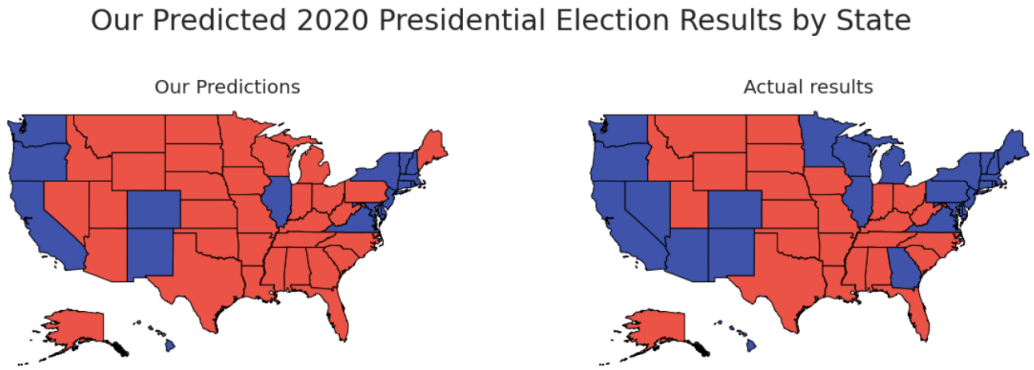

Figure 13: Our Predicted Presidential Outcome vs. Actual

Figure 13 shows our 2020 presidential predictions using our Ridge model and the actual results for each state. The ridge accuracy score was calculated to be 0.8431, while the accuracy score for our KNN model was 0.5098. As seen by the map comparisons, none of the red states in our actual results were misclassified as voting blue (while the opposite is not true). In light of this, we wanted to examine the false positive and false negative rates. Because 1 represented a Democratic win and 0 a Republican win, the false Democratic win rate was 0%, while the false Republican win rate was 24.24%. This confirms our visual analysis that our model successfully did not misclassify any of the red states as blue, so our model is conservative-leaning, likely stemming from our inclusion of lagged election outcome information as a predictor coming out of the 2016 Republican-victory.

Building upon our background knowledge of the U.S. political system and the concept of swing states, we noticed that our model was very successful at predicting non-battleground states (such as Alaska, California, New York), but struggled to predict the battleground states and the states that flipped from the 2016 election (Georgia, the midwest blue wall states, and Arizona). Thus, we wanted to explore the precision and recall of these two sets of states as define by the Cook Political Report [1].

Model Performance: Battleground vs. Non Battleground States			
Group	Precision	Recall	Accuracy Score
Battleground States	1.0	0.1111	0.4667
Non-Battleground States	1.0	1.0	1.0

Table 1: Table comparing results of Battleground vs. Non-Battleground states in the Presidential Election

Model Performance vs. 2020 Election Results			
Model	Accuracy Score	False Positive Rate (False Dem Prediction)	False Negative Rate (False Repub. Prediction)
Presidential Election	0.8431	0	0.2424
House Election	0.8683	0.104	0.166

Table 2: Table comparing Model Results

From Table 1, our model's strength in predicting the outcomes of non-battleground states is clear. Non-battleground states have been deemed by political scientists and the media to consistently vote Democrat or Republican. Stemming from our inclusion of lagged election outcomes as a predictor, our logit ridge model accurately predicted all of these. In battleground states, however, our model performed worse than random (0.5), with high precision, but low recall. In other words, our model classified most key battleground states as having Republican outcomes, though in this year's election where the flipping of swing states was a pivotal piece of Biden's victory, recall was key.

However, it's important to remember that 2020 was an unprecedented year and Donald Trump was a very unusual incumbent, who over the course of his presidency earned many atypical titles including being impeached in December of 2019. 2020 elections also featured unmatched voter turnout in a year where public health, police brutality, sky-high unemployment and climate change were brought to the forefront. In addition, nationally recognized forecasters such as The Economist and FiveThirtyEight both overly favored Joe Biden - FiveThirtyEight predicted North Carolina and Florida to be blue [6], while The Economist predicted an electoral vote of 356 to 182, when the actual electoral results were 306 to 232 [5].

It's also important to note that we did not differentiate between the multiple districts in both Maine and Nebraska - these states do not treat their results as a lump sum, but rather allot certain electoral votes to districts of their states. However, we treated these states as one unit.

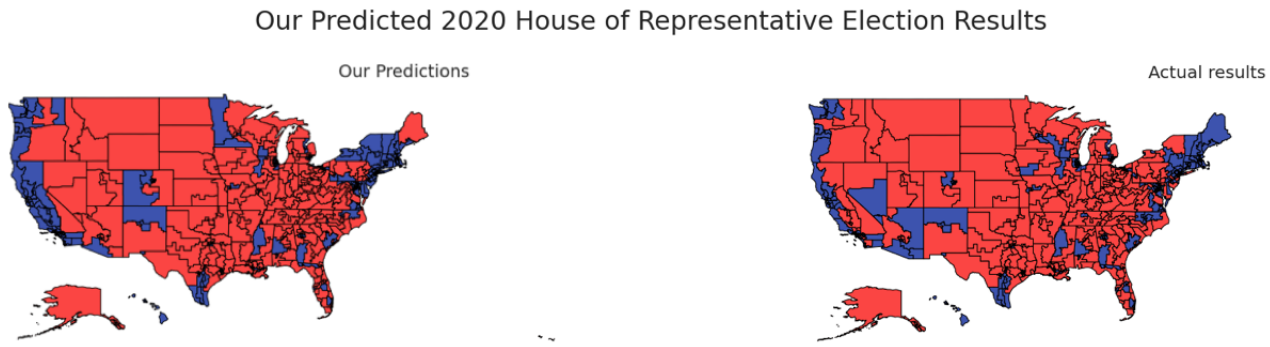

Figure 14: Our Predicted House of Representative Results vs. Actual

Figure 14 shows our 2020 House of Representative predictions using our Ridge model and the actual results for each district. The ridge accuracy score was calculated to be 0.8683. While our model performed well for predicting the House of Representatives races and exceeded the performance on the presidential election, we retain a conservative bias. In other words, the false negative rate exceeds the false positive rate,

as can be seen in Table 2. Again, these misclassifications mainly occurred in battleground states such as Nevada, Arizona, and Michigan.

6 Exploration of COVID Impact on Vote Outcomes

Finally, we were interested in exploring the relationship between voting outcomes and measure of impact from the COVID-19 pandemic. As shown in Figure 15, the onset of the pandemic was marked by a massive spike in unemployment [2] across the country, peaking in April but still lingering well above previous rates leading up to Election Day.

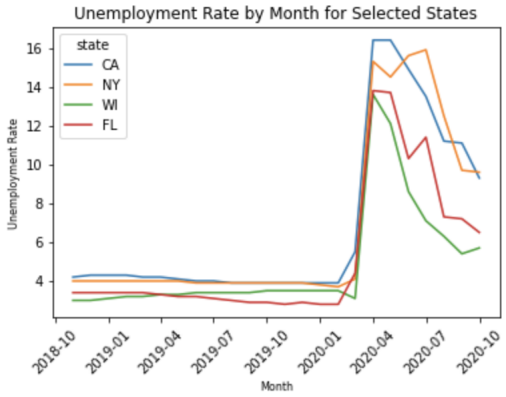

Figure 15: Monthly Unemployment Rate for Selected States

Looking at state unemployment leading up to past presidential elections, we noticed that unemployment rates have historically been far lower than they were this year. Unemployment was lower in 2016 than in 2012, though there seems to be some evidence of a positive relationship between unemployment rate and fold-change (where greater values indicate a higher-margin Democratic win and lower values indicate a higher-margin Republican win), as shown in Figure 16.

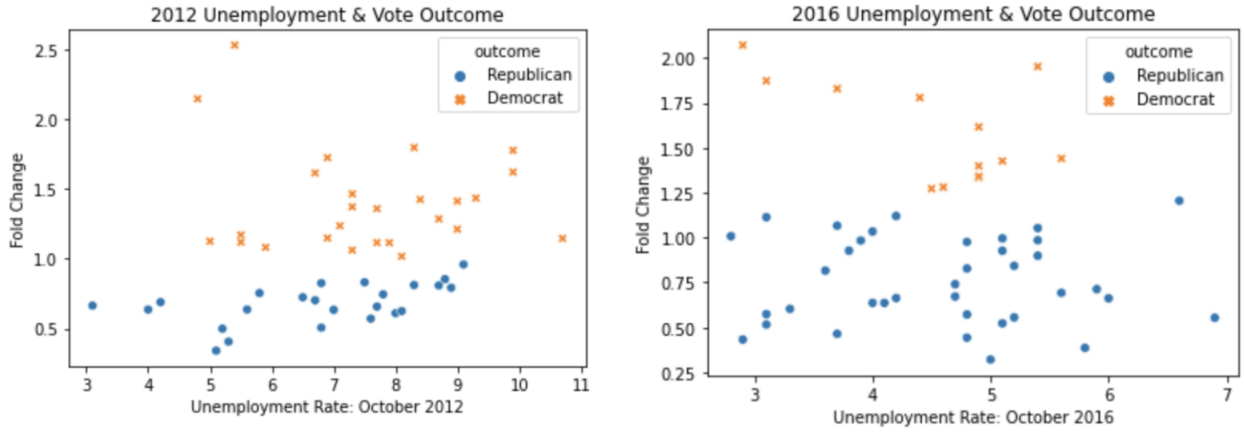

Figure 16: Historical Election-Year Unemployment & Fold Change

With this in mind, we calculated the increase in unemployment from January 2020 to October 2020 across states and plotted them in relation to our predicted and the actual state-wide binary election outcomes, as shown in Figure 17. Interestingly, it appears that predicted and actual Democrat-won states had higher

increases in unemployment. However, it's worth noting that since the data is aggregated at the state level, it may be more meaningful to evaluate unemployment rates on a more granular level, such as by county or Congressional district. It could be the case that states with population-dense urban centers and industries (e.g. restaurants) that were more heavily impacted by COVID are also those that tended to vote blue.

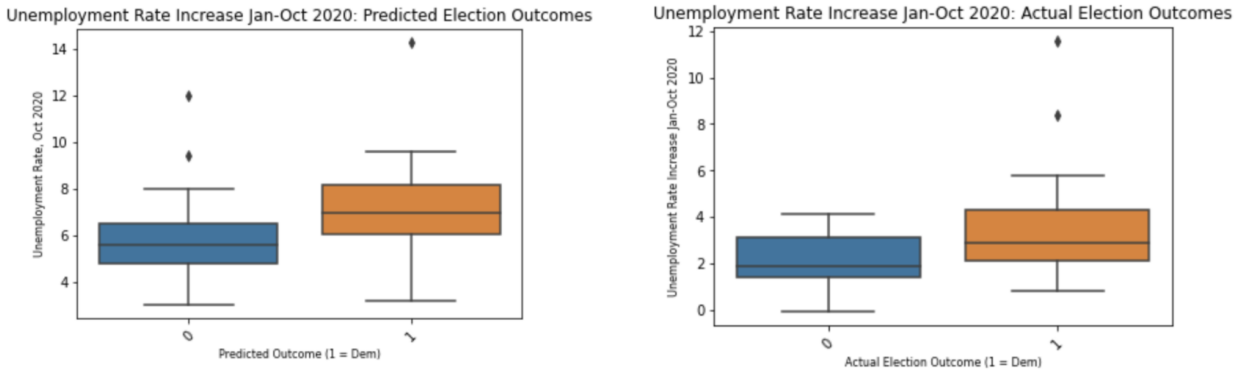

Figure 17: Increase in Unemployment vs. Election Outcome (Pred & Actual)

We also looked at voting outcomes in relation to direct measures of COVID prevalence. Using data on state COVID case rates over the past week[4], we plotted the number of COVID cases per 100 thousand people (over the past seven days from when the data was pulled on December 10). Based on the plots shown in Figure 18, it seems to be the case that predicted and actual Republican states have, on average, greater rates of COVID per-capita. Naturally, the pandemic and its grip on society has had a substantial impact on life across America. It is fascinating to consider the relationship between pandemic policy responses and election trends.

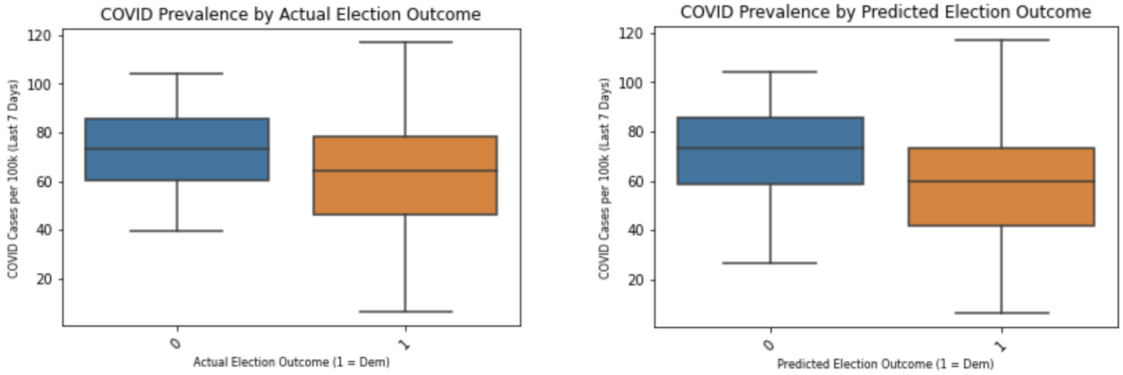

Figure 18: COVID Prevalence by Actual & Predicted Election Outcome

7 Social Impact Statement

Our presidential model is biased towards predicting Republican wins, as evidenced by the disparate False Democratic Predictions vs. False Republican Predictions in Table 2. Reflecting on Elís Miller Larsen's lecture on responsible data science, our model is thus not fair. This is a learning problem with real world impact and needs to be both accurate and fair, so it is vital for us to continue iterating before putting this into production.

Through this project, we learned how difficult of a learning task predicting election results is - ultimately, we are trying to predict the psychology of a nation of over 300 million people during unprecedented times. Fundamentally, this provides more justification for why it is important that we don't rely on election forecasts[3]. Forecasters' reliance on opinion polls of unrepresentative samples of the population have notoriously resulted in inaccurate predictions.

When forecasts are incorrect or when we rely on these forecasts too heavily, there can be major consequences. Potential voters are discouraged from participating in their civic duties because of an underlying belief that the election was already decided, we don't have an accurate understanding of what voters want, and we underestimate the prevalence and importance of minority votes. Another challenge of this year in particular is a lack of trust in the democratic process [7]. We should not be relying on certain forecasts or certain people's intuition as to what the election results should be, rather we need to listen to state election officials, trust their certified votes, and not succumb to disinformation. American elections are powerful, and the world pays attention. It is too big of a stage for any person's or any organization's predictions to carry too much weight.

8 Video Presentation

YouTube Link: https://youtu.be/UlOpDCXd2mA

References

[1] *2020 Cook Political Report.*

[2] Bureau of labor statistics state & local unemployment rates.

[3] Can we finally agree to ignore election forecasts?

[4] Cdc covid data tracker. $https://covid.cdc.gov/covid\text{-}data\text{-}tracker/cases_{c}asesper100klast7days$.

[5] The economist- forecasting the u.s. elections.

[6] Fivethirtyeight elections forecast.

[7] Supreme court shuts door on texas suit seeking to overturn election.

[8] Andrew Mercer, Claudia Deane, and Kyley McGeeney. Why 2016 election polls missed their mark. *Pew Research Center*, Aug 2020.