Predicting U.S. Election Results in Swing States

Vaed Prasad (vsp22), Emily Weed (emw232), Alexander Salonga (as2739)

December 2020

1 Introduction

Unlike democratic elections in most nations where leaders are elected based on the popular vote of its constituents, presidential elections in the United States follow a unique electoral system outlined by the Electoral College. In this process each state is allocated a share of electoral votes that is roughly proportional to its population and, in predominantly every state, the winner of the statewide plurality collects all the electoral votes from that state. This structure presents a phenomena where the outcome of an entire presidential election can be decided on a handful of swing states, where either presidential candidate has a feasible opportunity of winning the statewide plurality. In this project we aim to leverage historical demographic, socioeconomic, and electoral data to predict which candidate will win these critical swing states in the 2020 United States Presidential Election. For additional details regarding our goal, please reference our proposal.

2 Exploratory Data Analysis for the 2016 Election

2.1 Data Selection

In order to predict the results of the 2020 U.S. Presidential Election, we honed our focus to twelve key swing states: Michigan, Florida, Nevada, Texas, Minnesota, Wisconsin, Iowa, Ohio, North Carolina, Georgia, Arizona, and Pennsylvania. These states were selected by considering the states with the closest margin in the 2016 U.S. Presidential Election. For these twelve states we collected county level data containing racial composition, age and gender clusters, economic features, health conditions, occupation, educational level attained, etc. We plan to use this data to analyze their respective influence on the county's voting distribution for the Republican and Democratic presidential candidates in the 2016 election.

2.2 Data Cleaning

In Data Cleaning and Manipulating we compiled our various data sets into a single dataframe where each row represented a particular U.S. county. In order to facilitate the merging of various datasets, we utilized the "Fips" feature, a unique identifier for all counties in the United States, as a primary key.

Several features containing numerical data in the CSVs were encoded with data type String; we cleaned this features by converting them to type Int in order to facilitate running regressions and measuring errors for our model.

One consideration we had to make when cleaning our data was that a number of features were represented as a raw count (i.e. "AGE1824-MALE", "AGE1824-TOT"), which would be heavily dependent on the county's population. As a result, we normalized these particular features by dividing their raw count by the "Total Population" column, so that we could consider a more relevant per capita representation of these features. Next, we pruned features with a large density of missing values by calculating the null ratios for each feature and dropping features that had a null ratio greater than 0.05. Thinking ahead to collecting data for 2020, many of these columns were very specific measures and would be difficult to find reliable updated values. We decided it would to drop these columns as they had a large percentage of missing values and would not add value to our model even if we imputed the missing values.

2.3 Feature Selection

After cleaning our data we had a large slate of merged, normalized, and filled data to consider for our final feature selection. In order to pick the most relevant features for our model we constructed a correlation plot to observe a feature's relationship with the Democrat/Republican vote share. We were aiming to select about 20 features (besides the 2012 vote share) to work with further so we chose the subset of 20 features with the largest | Correlation of x with "Democrat 2016" or "Republican 2016" |. We constructed the correlation plot for our final subset of features. See Figure 1 below.

2.4 Feature Transformations

We created scatter plots of our features against our target variables to assess the linearity of the relationships. We observed that every feature except "realGDP2016" showed a linear relationship with the target variables.

Figure 2 displays the distribution of points before performing any feature transformations. From Figure 2 one can deduct that because there is such a high density of counties with small GDPs with varying levels of Democratic vote share, it is challenging to identify any apparent trend. In order to mitigate this clustering dilemma we calculated the log_{10} of "realGDP2016." As observed in Figure 3, this transformation creates a more identifiable linear relationship between the variables, where an increase in log_{10}realGDP2016 tends to result in an increase in Democratic vote share. All considered features for our model are continuous numerical variables, so we did not have to encode any nominal, ordinal or text features.

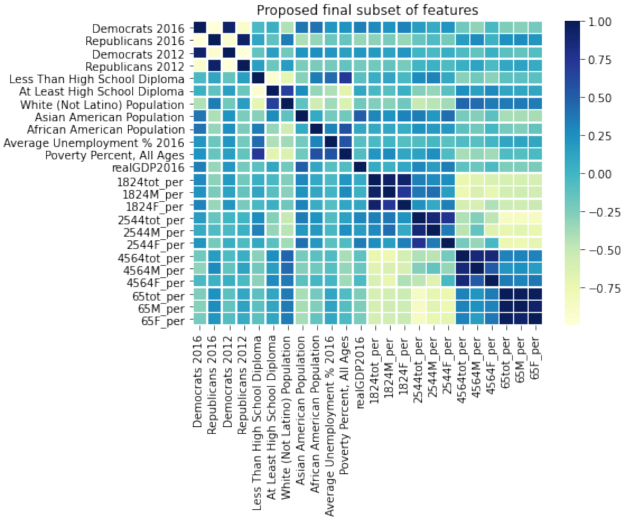

Figure 1: Correlation Plot of Final Feature Selection

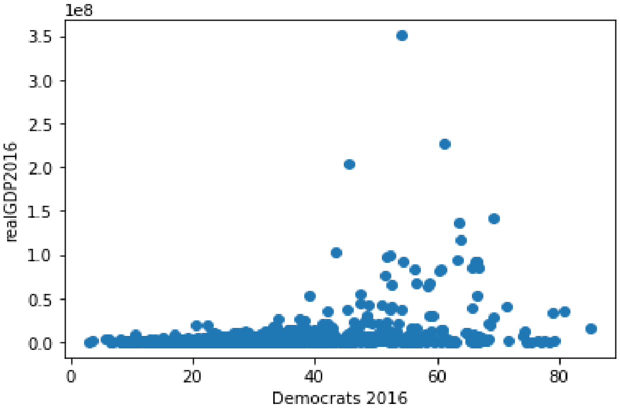

Figure 2: County Level Real GDP vs. Democrat Vote Share

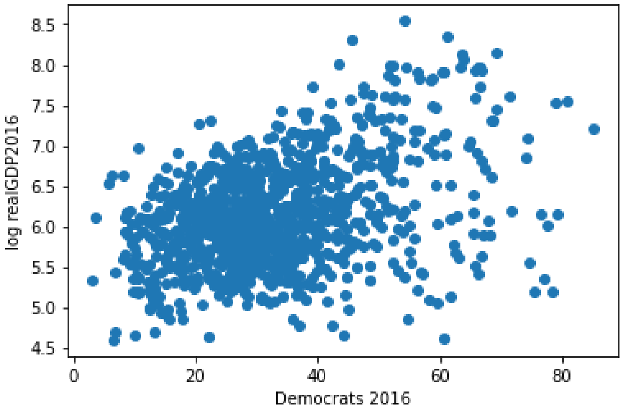

Figure 3: Log County Level Real GDP vs. Democrat Vote Share

3 Modeling the 2016 Election

3.1 Model Construction

We plan to use 5-fold cross validation to augment our understanding of the test error before performing any predictions with our test set. This will also improve our sense of whether our model is overfitting or underfitting by making better estimates of the test error to adjust our model.

We are using a Linear Regression Model. Initially, we attempted to observe how well our model performed without using previous year's data on Democrat/Republican vote share. We suspected that this would not have a very high accuracy but it serves as a baseline for our future models. In Model Construction we got training and validation R^2 around 0.55, train MSE (mean squared error) and validation MSE both around 75. The R^2 demonstrates the amount of variation in the response variable the features explain, where an R^2 closer to 1 indicates that there is a high proportion of variance in the vote share that is predictable from our features. Although the MSEs are rather large for our initial objective, we can observe that our model is neither overfitting nor underfitting because the training and validation R^2 and MSE values are in proximity.

Next, we incorporated the vote share from the preceding presidential election, which provided strong insight on the likely outcomes of the 2016 election. As hypothesized, our R^2 and MSE values greatly improved to 0.978 and 4.5 respectively, for both training and validation. We also isolated the previous election year's vote share to understand the influence of our other features. This model performed better than our initial model, without vote share from the previous presidential election; however, this model did not perform as well as the model that included our entire slate of features. We concluded that these results indicate that we should further investigate leveraging a combination of all the features and that

Variable	Type
Democrats 2016	Continuous
Republicans 2016	Continuous
Democrats 2012	Continuous
Republicans 2012	Continuous
No HS Diploma	Continuous
Least HS Diploma	Continuous
White	Continuous
Asian American	Continuous
African American	Continuous
Avg. Unemployment	Continuous
Poverty Percent	Continuous
Real GDP	Continuous
18-24 Total	Continuous
18-24 Male	Continuous
18-24 Female	Continuous
25-44 Total	Continuous
25-44 Male	Continuous
25-44 Female	Continuous
45-64 Total	Continuous
45-64 Male	Continuous
45-64 Female	Continuous
65+ Total	Continuous
65+ Male	Continuous
65+ Female	Continuous

Table 1: Final Feature Selection

the demographic and economic features have a significant contribution in improving our model's accuracy.

The subset of features we decided to proceed with is enumerated in Table 1, excluding the gendered breakdown of age groups. Training a model on these features yielded a similar training and validation R^2 and MSE to our full model with all considered features. Figure 4 presents a scatter plot of the predicted Democrat vote share against the actual Democrat vote share, where each point represents a county. Counties on par with our linear regression model, represented by the dotted line, are colored dark blue while large outliers are colored light yellow. Figure 5 illustrates a histogram of the difference of the predicted Democrat vote share from the actual Democrat vote share ($Actual - Predicted$) for each county in the test set. The performance between the model with all features was nearly identical to the model with the subset of features. As a result, we selected the model using the pithy subset of features as sparse models are easier to interpret and can reduce the number of necessary observed examples to learn the model.

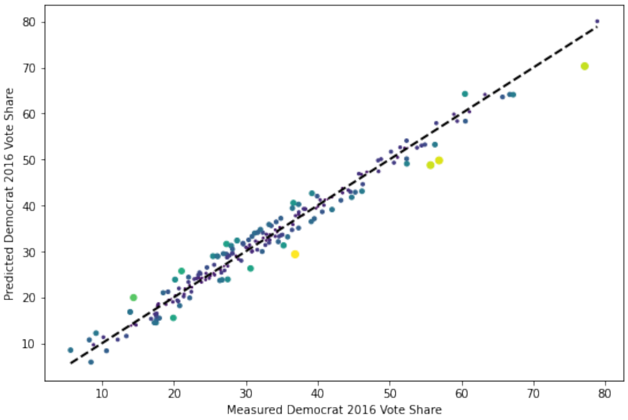

Figure 4: Scatter Plot of Predicted Vote Share vs Actual Vote Share

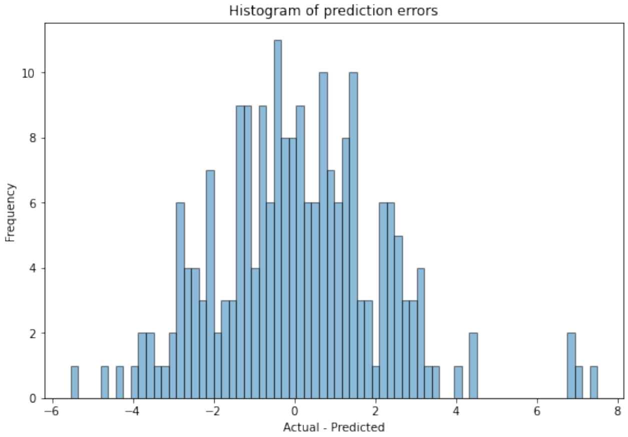

Figure 5: Histogram of Actual - Predicted Vote Share

4 Exploratory Data Analysis for the 2020 Election

4.1 Data Selection

In order to take advantage of our model fitted on the 2016 data to predict the outcome of the 2020 Presidential Election, we needed to collect equivalent data in 2020 for all the features in our model.

We were able find the majority of the same variables, from either 2019 or 2020 depending on the most recent release dates, for the data. In addition, we were able to find recent county-wide unemployment rates from October 2020, which we believe will help capture the economic effects of the pandemic on these counties. Unfortunately, we were unable to identify a complete data set containing recent education and race data for all of our counties. The only recent public data for these variables were in the American Community Survey from the Census; however, it only reported values for a portion of the counties in each state. The report left out a great deal of the smaller counties in the swing states we were analyzing which bottle-necked our data set size, reducing the number of complete swing state county rows from 1064 to 325. We decided to proceed with this data set as one of our approaches with the constraint of missing many counties because this data is the most recent, reliable we could find. Going forward, we will refer to this data set as our small 2020 data set.

For our second approach, in order to preserve the continuity of testing our model on the same set of counties as our original 2016 model was trained on, we expanded our search and were able to find all the variables except for education (Less than High School, At Least High School, At Least Bachelor's Degree, and Graduate Degree) for our original full set of counties. In order to use this data set without the education data, we had to retrain our 2016 model, dropping the education data from that data set. Going forward, we will refer to this data set as our large 2020 data set.

4.2 Data Cleaning

We preformed many of the same data cleaning processes on these data sets as we did with out original 2016 one. We had to merge numerous data sets together, utilizing the 'Fips' identifier if it was available. We again calculated the log_{10} of "realGDP2019" to maintain consistency with our 2016 model and produce a linear relationship between real GDP and vote share. We realized we had to transform the vote share to be on the same scale as the vote share was in the 2016 data set, e.g. transforming 0.436 to 43.6. Additionally we had some missing values in the races variables (White, Asian American and African American) in our smaller 2020 data set. To remedy this we used the Iterative Imputer library in python. This library helps impute missing data by modeling each feature as a function of the other features. The columns with missing values are imputed sequentially, using a Bayesian Ridge Regression model to predict the values. Bayesian Ridge Regression, a form of regularized linear regression, uses probability distributions to predict the response variables (the missing values in our case) and adapts to the data available resulting in imputations that preform better than using the mean or median.

5 Running Our Models on 2020 Data

5.1 Using Our Small 2020 Data Set

From here on out, our target variable is the Democrat vote share in the 2020 election as our 2016 model was trained using the Democrat vote share in 2016 as the target variable in this case. First, we ran our models that were trained on 2016 data on the smaller data set we gathered for 2020 (325 counties). There were three models we developed: Model 1, trained with all the variables, Model 2, trained with only the previous election year's vote share as the features, and Model 3, trained with all the features except the gendered breakdown of population. See 3.1 for more information.

We initially ran Model 1 on our collected 2020 data set. We got an MSE value of 14.43 (compared to a MSE of 4.25 when run on the 2016 test set). From this discrepancy in the MSE values we can begin to notice that our model is not preforming particularly well on the 2020 data. Next, running our model 2 on the small 2020 data, the gap between the MSE values grows. For the 2020 data it is 329.40, and for the 2016 test data it is 23.28. We knew this model did not perform as well beforehand due to the estimation of the test error we had from our 5-fold cross validation approach in 3.1. Finally we ran model 3, our main model, on the small 2020 data set. This gave a MSE of 49.44 (MSE of 3.90 when run on the 2016 test data). This also demonstrates a large gap in the performance of the model between the 2016 and the 2020 data.

Figure 6 shows a histogram of the actual 2020 vote share - our predictions from model 3. The distribution appears to be almost normal however it is not centered around 0 as we hoped. The median appears to be around 6 or 7 which indicates that our model was consistently under predicting the Democrat vote share. In Figure 7 we plotted a scatter plot of the predicted vote share vs. the actual vote share, with a dotted line representing perfect predictions. This scatter plot deviates from the dotted line almost immediately. This figure shows that as the actual Democratic vote share increased in the county, the model prediction was worse, specifically under predicted more significantly.

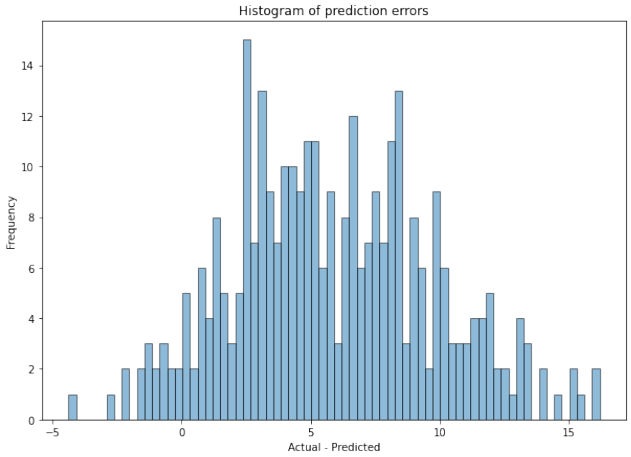

Figure 6: Histogram of Actual - Predicted for 2020 Small Data Set

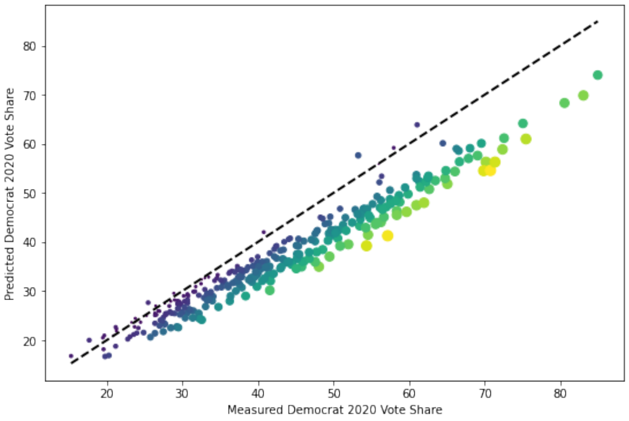

Figure 7: Scatter Plot of Predicted vs. Actual for 2020 Small Data Set

5.2 Using our Large 2020 Data Set

Next we proceeded with our data we collected for 2020 that included all the counties we originally had for 2016. In order to use this set, we had to sacrifice the use of the Education variables. We were unable to find reliable data for these values for 2019 or 2020 time frames. Due to this, we had to retrain our 2016 models without the use of the education features. We went through the process of dropping those from the 2016 data set, and constructing new models trained on 2016 data. We trained one using all the features available (model 4 going forward) and one with all the features available except the gendered breakdown of age (model 5).

Running model 4 on our 2020 data gave us a MSE value of 93.32 (compared to a MSE of 7.86 when run on the 2016 test data). Similar to our work with the smaller data set, this is a big difference in MSE values. Running model 5 gave us similar values for a 2020 MSE and 2016 test MSE, 90.24 and 7.90 respectively. This is, again, a big difference.

We made the same visualizations as we did for the small 2020 data set in order to asses how the predictions were distributed. The histogram in Figure 8 mirrors that of Figure 6; the values seem normally distributed but are not centered around 0, the model again under predicts the vote share. In this graph the median appears to fall around 9 which is higher than in Figure 6. This suggests that using more counties at the expense of the education variables resulted in a slight decrease in performance of our models on 2020 data. Figure 9 shows the same scatter plot as previously described and once again we can see that the model under predicted the vote share consistently. This graph differs from Figure 7 as there does not appear to be a downward trend in the points. Figure 9 indicates it was consistently poorly predicting the vote share regardless of the true value.

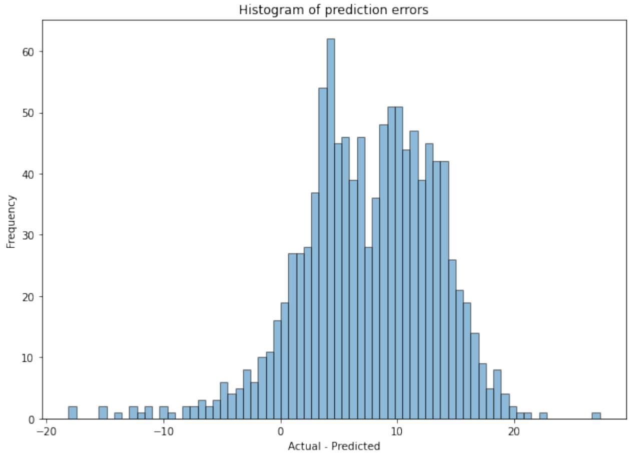

Figure 8: Histogram of Actual - Predicted Vote Share for 2020 Full Data Set

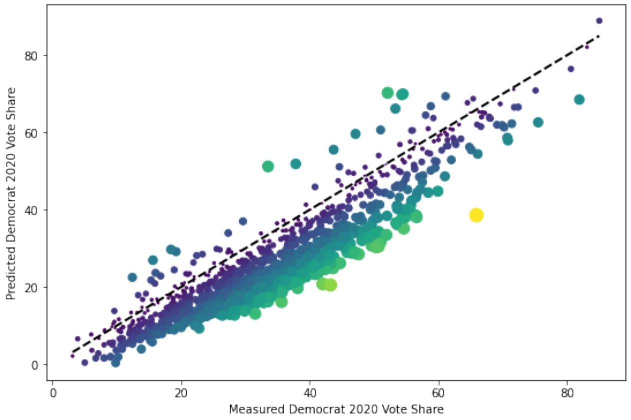

Figure 9: Scatter Plot of Predicted vs. Actual for 2020 Full Data Set

6 Assessing our Predictions

In order to quantify our predictions in terms of determining who our model predicted would win the 2020 presidential election we decided to compute a weighted average of the predicted vote share across the counties. This would then give us an estimate of the statewide percentage for the Democratic candidate. Some limitations of this approach we acknowledge are that we did not have data for the exact number of registered voters per county prior to the 2020 election. We instead used the total population 18 years and up per county as our normalizing factor. We did this for each of our swing states and compiled the results into arrays we then graphed in a similar scatter plot as before. We computed these averages using our predictions from our small 2020 data set, as illustrated in Figure 10, and our full 2020 data set, as illustrated in Figure 11. These both show that our state wide predictions, done in this matter, were off by a great deal. While, we acknowledge that this method for calculating the state wide vote share, and thus who wins that state, is not the most accurate, the main problem here is our actual county level predictions.

Since our predictions on both our smaller 2020 data set and the full 2020 data set were both consistently under the true value, we can hypothesize that the problem is with our original model construction. It seems that it was not able to capture the blue shift many states and counties saw in 2020. This can be due to a number of factors. The most obvious of such would be the pandemic and the current administration's response. We decided to not directly account for this in our model as it is difficult to quantify this idea. We thought that the unemployment rates, since it was data from October 2020, would help in addressing the major changes counties have seen since the pandemic began but it seems it was not enough. Additionally there are countless factors that are very difficult to quantify that influence someone's choice on who to vote for.

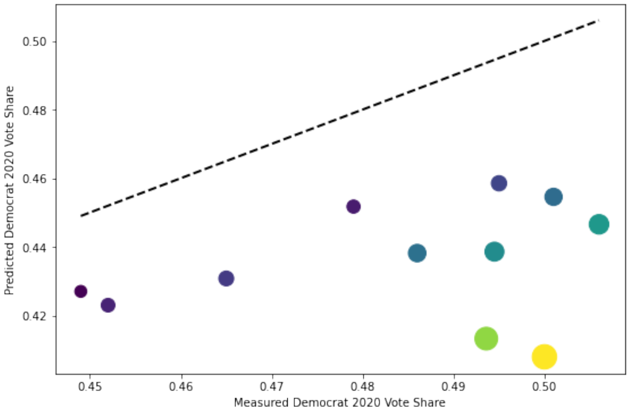

Figure 10: Scatter Plot of Predicted vs. Actual for 2020 State Wide Vote Share from Small 2020 Data Set

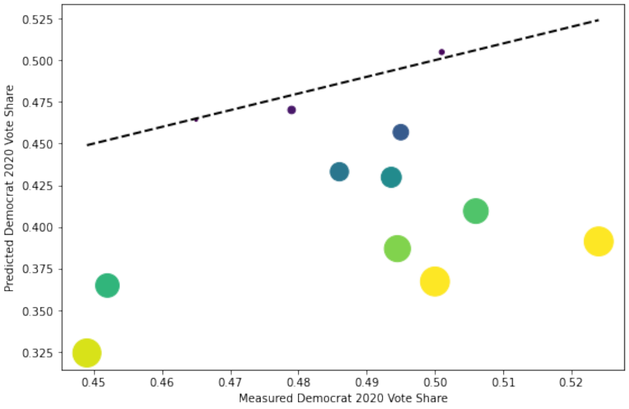

Figure 11: Scatter Plot of Predicted vs. Actual for 2020 State Wide Vote Share from Full 2020 Data Set

7 Mail In Voting

We wanted to explore the topic of voting by mail since it was so widely used in the 2020 election. We knew beforehand that many more people would vote absentee for 2020 because of the COVID-19 pandemic. We also wanted to explore the party breakdown for the 2020 absentee votes. Collecting reliable data for these measures was a little difficult as some states do not report party registration for mail in ballots. We were able to find some sources and compiled them into visualizations.

Figure 12 shows the turnout for the general elections from 2000 up to 2020. We included the swing states we focused on for the modeling portion of this project. We can see that some states saw a major jump in turnout for the 2020 election (Florida, Texas, Arizona) while others seemed to increase consistently with the previous years changes.

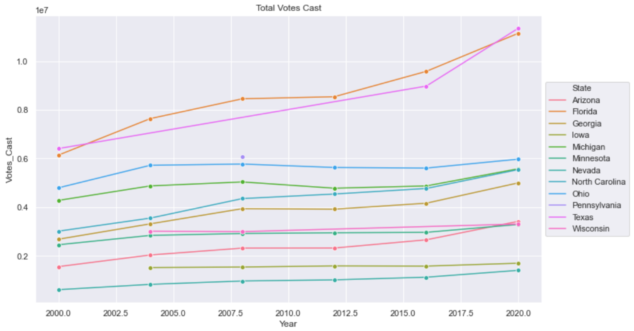

Figure 12: Total Ballots Cast 2000-2020

Figure 13 shows the absentee ballots returned for our 12 swing states broken down by year. Some states do not report values for this metric (or haven't released reliable statistics for 2020) hence the lack of bars for some states. Regardless, we can clearly see that in 2020 many more people chose to vote by absentee ballot. It looks to be double in most states. This data supports the statement that more people voted by mail this year than in previous years as was expected.

Figure 14 shows the number of absentee ballots returned for 2020 broken down by party registration. In this graph we wanted to explore the statement 'Democrats were more likely to vote by mail in the 2020 election'. We only included states which reported values for this metric in the plot to make it easier to read, as some states do not report party registration at all. Overall we can see a trend that supports this statement. In all states except Arizona, the difference in the heights of the bars is considerable, with Democrats being higher. Pennsylvania has the biggest difference which was expected as mail in ballots were the deciding factor to swing Pennsylvania blue in 2020.

In conclusion, these graphs give good baseline insight into absentee voting. Investigating the changes in mail in voting and overall voter turnout over the years has

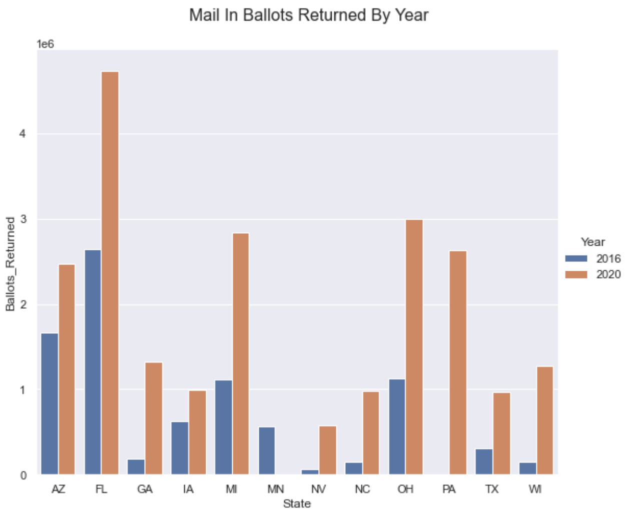

Figure 13: Mail in Ballots Returned 2016 and 2020

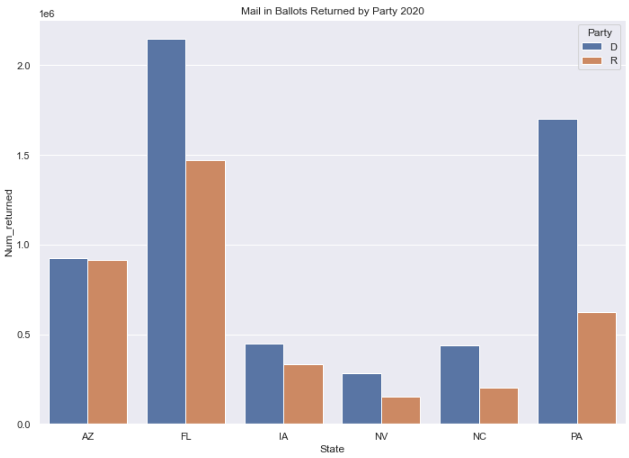

Figure 14: Mail in Ballots Returned 2020 by Party Registration

the potential to be its own project entirely. We just wanted to explore it on the surface to support our model results. These graphs do a good job of providing basic interpretable results.

8 Limitations of Data Science

8.1 Weapons of Math Destruction

Data Scientist Cathy O'Neil, author of *Weapons of Math Destruction*, coined the term to describe models that shield themselves as neutral mathematical algorithms but actually distort the truth and reinforce inequality.

If we had a highly accurate model to see which factors can significantly impact the outcome of the election, it would have a profound impact on voter turnout. If a model could accurately predict the outcome of the U.S. presidential election before the first ballot is even cast, voters may be deterred from participating in the democratic process as their vote will appear moot. Similarly, the will of certain minority groups may be considered insignificant since their electoral vote share is negligible. This may in turn cause a feedback loop where the candidate focuses on a certain group causing the model to show that this group has an even stronger impact on the vote share.

On a similar note, consider how U.S. presidential candidates campaign. The election is usually determined by a few key swing states and as a result candidates focus their campaign resources on them. Non-swing states which compose the majority of the country are largely ignored by the candidates during the few weeks leading up to election day.

Beyond influencing campaign strategies, election predicting models can enable politicians to essentially gerrymander districts across other features. Current politicians may be incentivized to pass legislation furthering their party's vote share rather than representing their constituents. In addition, politicians can also be motivated to discriminate against groups that are unfavorable for the candidate. These effects can have a detrimental impact on the integrity of our election system.

8.2 Fairness

In terms of fairness, our model definitely has the potential for algorithmic bias. The immense amount of data that needs to be collected in order to create a reliable model may not be readily available for minority populations. This can cause the model to be biased towards the groups with more reliable statistics.

Although The Equal Housing Opportunity Act and The Equal Credit Opportunity Act regulate protected groups from algorithmic bias in the housing and banking industries, no such regulation exists for electoral campaigning. As a result, candidates are encouraged to pander to citizens' demographics rather than recognize their individuality.

9 Conclusion

Although we were able to construct a model that performed well within the trained election year, when we attempted to extrapolate our model for the 2020 U.S. presidential election, the accuracy did not hold. Election predicting is a nuanced problem as there are a myriad of immeasurable features that influences who an individual votes for. We ran into issues collecting reputable data for 2020 and attempted to remedy this by expanding our search. We were able to recognize that our model pretty consistently under predicted the Democrat vote share across the board.

We also recognized the stark contrast between the participation of the 2016 election and the 2020 election. While voter turnout in the 2016 election was on par with previous elections, according to the Washington Post the voter turnout in the 2020 election was the highest in over a century at a staggering 66.2%. As a result, our model fitted on the 2016 election was not able to accurately predict the "Blue Wave" of votes for the Democratic candidate Joe Biden. However, for the demographic, socioeconomic, and electoral features that we were able to measure, we observed that they did have a substantial effect on a county's vote share. We gained insight at how these features influence an individual's decision to vote Democrat or Republican in presidential elections.

10 References

- County Level 2016 Electoral Results
- 2016 County Level Age and Gender Distributions
- County Level Economic Data
- 2020 Mail In Voting
- 2016 Mail In Voting
- 2020 Race Data
- 2020 Education Data
- 2020 Poverty Data
- 2020 Election Results