# ”That’s All, Votes!”: Factors Influencing Presidential Vote Outcomes at the County Level in the 2016 U.S. Presidential Election

Alexander Adams

December 19, 2020

12 Pages, 3,934 Words

## 1 Abstract

The aim of this project is to assess the effectiveness of various factors in predicting which of the two major party candidates (Democrat Hillary Clinton and Republican Donald Trump) won more votes in a given county in the 2016 United States presidential election. The purpose of this report is to discuss the motivations for this project, the data gathering and cleaning process, analyze and present the results of the most effective machine learning methods, and consider possible next steps for this research topic. This project also attempts to draw conclusions from misclassified observations data, and briefly discusses potential real-world implications of the machine learning results.

## 2 Problem Statement and Background

The outcome of the 2016 presidential election came as a surprise to a large number of voters and observers in the United States and around the world. Many (myself included) expected Hillary Clinton to defeat Donald Trump, a prediction based off of polling and forecasting from various media outlets. While Clinton won the popular vote, Trump won the electoral college, and thus the presidency. After the election, political analysts, particularly Democrats, felt it necessary to figure out what went wrong, and to learn from 2016. The focus of this project can be stated as follows: 1) what factors predict which major-party candidate received more votes in a given county in the 2016 U.S. presidential election, and 2) are there any notable patterns present among incorrectly predicted counties?

## 3 Literature Review

Popular discourses and punditry around voting in the United States are predicated on a number of heuristics relating to socioeconomic and demographic factors, including

the effects of race, partisan affiliation, economic indicators such as unemployment, education, and the increasing rural-urban divide. Some of these predictors were accurate in 2016, while others were not; regardless, all merit inclusion in this analysis. Regarding race, nonwhite voters of all races overwhelmingly vote for Democratic candidates[1]; while white voters are more divided, the Republican candidate has historically won a plurality or majority of white votes in modern American elections. In previous elections, nonwhite voters turned out at lower rates than white voters, with the exception of 2012, when Black voters cast ballots at a higher rate than white voters. Nonwhite voter turnout, particularly among Black voters, increased from 2004 to 2008 and 2008 to 2012. However, the gap between white and nonwhite voter turnout increased from 10% in 2012 to 12.6% in 2016, with Black voters turning out at a lower rate than any election since 2000.[2] As such, it is reasonable to expect that counties with larger nonwhite populations are more likely to vote for a Democratic candidate, though this effect may have been diminished somewhat in 2016 due to decreases in Black turnout.

In addition to racial divides, the United States is increasingly experiencing a divide between rural and urban areas, and between Americans with different levels of education. Urban centers and major metropolitan areas consistently support Democratic candidates, while rural communities tend to vote Republican, a trend which exists regardless of region and, like partisan sorting, has been progressing for the past several decades.[3] Education (or rather, the lack of weighting by education) is also frequently cited as a failure of election polling in 2016 (particularly at the state level). A review of one state-level poll of the 2016 election conducted by the University of New Hampshire found that weighting responses by education alone was enough to remove virtually all of the error in the polling results.[4] Another potential division occurs along the dimension of economic health. Many voters practice economic voting, or choosing who to vote for based on economic issues. The literature on economic voting, specifically as it pertains to unemployment, is inconclusive; some of the literature suggests that the Democratic party benefits when unemployment is high, since that party is perceived as being stronger on economic issues, while other research suggests a more general negative effect on the incumbent party.[5] In 2016, the Democratic party was the incumbent party, so it is unclear if unemployment produces an effect on vote choice. Unlike the other heuristics discussed above, the uncertainty surrounding the effect of unemployment on Democratic victories in a year where the Democratic party is the

1. While there exist differences between nationalities within broader racial categories (such as Cuban Americans versus Mexican Americans, or Vietnamese Americans versus Chinese Americans), such differences are beyond the scope of this inquiry.

2. William H. Frey, "Census shows pervasive decline in 2016 minority voter turnout," December 18, 2020, accessed December 18, 2020, https://www.brookings.edu/blog/the-avenue/2017/05/18/census-shows-pervasive-decline-in-2016-minority-voter-turnout/.

3. Seth C. McKee, "Rural Voters and the Polarization of American Presidential Elections," *PS: Political Science and Politics* 41, no. 1 (2008): 101–108, ISSN: 10490965, 15375935, http://www.jstor.org/stable/20452117.

4. Courtney Kennedy et al., "An Evaluation of the 2016 Election Polls in the United States," *Public Opinion Quarterly* 82, no. 1 (February 2018): 1–33, ISSN: 0033-362X, https://doi.org/10.1093/poq/nfx047, eprint: https://academic.oup.com/poq/article-pdf/82/1/1/24265180/nfx047.pdf, https://doi.org/10.1093/poq/nfx047.

5. Taeyong Park and Andrew Reeves, "Local Unemployment and Voting for President: Uncovering Causal Mechanisms," September 28, 2018, accessed December 18, 2020, https://andrewreeves.org/papers/unemployment.pdf.

incumbent presidential party means that it is not possible to predict what, if any effect unemployment rates will have on the probability of a Democratic victory.

While the aforementioned variables are important, the impact of partisan affiliation potentially outweighs them all. The past few decades have seen the process of partisan sorting[6] nearly completed.[7] Based on national-level polling, Clinton and Trump each won nearly 90% of self-identified partisans within their respective parties. To invert the old adage, past performance may in fact predict future results.

# 4 Data Sources and Processing

The data set used in this project consists of 3,139 observations of 114 variables (1 outcome variable and 113 feature variables). The unit of observation for this data set is the county or county-equivalent level. I specify county-equivalent because Alaska is divided into boroughs and borough-equivalent census areas, Louisiana is divided into parishes, and some cities are formally incorporated as independent cities and are considered equivalent to counties. Each county or county-equivalent has a unique five-digit Federal Information Processing Standards (FIPS) code. The 50 states and the District of Columbia collectively encompass 3,142 FIPS codes; the three codes not included in the data are all in Alaska, and were dropped due to high levels of missingness (including the dependent variable).

One issue with this research question arises from the way the United States administers elections. The combination of winner-take-all elections (with the exceptions of Nebraska and Maine) and the electoral college system means that votes are effectively allocated at the state level. Because of this, county level data is thus limited in the information it can provide about winning elections (since someone can win most of the counties in a state but still lose the state, depending on how populations are distributed). However, a county-level data set can contain over 3,000 observations, while a state level data set can only contain 51 (including the District of Columbia). Even using state-year as the unit of analysis and pooling data from election cycles going back to 1976 would only result in fewer than 20% of the observations in a county-level data set. This small data set would then be split into training and test data, meaning the model would be trained on an even smaller number of observations. I would expect any model used on state-level data to overfit the data and have limited predictive power on the test data.

I gathered my data from several sources. The data on unemployment came from the U.S. Department of Agriculture.[8] The data on demographics (race, gender, and age) came from the U.S. Census Bureau.[9] The data on vote totals and outcomes came from the MIT Election Lab.[10] The data on vote totals and outcomes for the state of Alaska

6. i.e. liberal partisans aligning with the Democratic party and conservative partisans aligning with the Republican party

7. Morris P. Fiorina, "The Political Parties Have Sorted," in *Essays on Contemporary American Politics* (Stanford CA: Hoover Institution Press, 2016), 1–20.

8. U.S. Department of Agriculture, *USDA ERS - Download Data*, October 30, 2020, accessed October 30, 2020, https://www.ers.usda.gov/data-products/county-level-data-sets/download-data/.

9. U.S. Census Bureau, *County Population by Characteristics: 2010-2019*, December 17, 2020, accessed December 17, 2020, https://www.census.gov/data/tables/time-series/demo/popest/2010s-counties-detail.html.

10. MIT Election Data and Science Lab, *U.S. President 1976–2016*, v. V5, 2017, https://doi.org/10.7910/

came from rrhelections.com, an explicitly Republican elections website.[11] While using data from a partisan source is not ideal, I have not observed any quality issues in the data.[12] The data on marriages and divorces was gathered by researchers at Bowling Green State University,[13] and the data on health and social indicators was gathered by researchers at the County Health Rankings Project run by the University of Wisconsin Population Health Institute.[14]

Overall, the data used for this project did not demonstrate a significant degree of missingness (no variable had more than 2-3% of its observations missing). The dependent variable for this analysis is a dichotomous variable which codes whether Clinton or Trump received more votes in a county in 2016. This variable is coded as 1 if Clinton won more votes, and 0 if Trump won more votes. Given that the overwhelming majority of votes cast in 2016 were for one of the two major-party candidates, I elected to ignore third parties in this analysis.

The 113 feature variables used for this project can be broken down as follows (all variables are solely for 2016 unless otherwise noted):

- Unemployment rates and county workforce size for each year from 2000 to 2016, inclusive
- Rural-Urban Continuum Code, Urban Influence Code, and Metropolitan Designation (2013)
- Proportion of the two-party vote share received by the Democratic candidate in election cycles from 2000 to 2012, inclusive
- Total number of votes cast for all candidates regardless of party in election cycles from 2000 to 2012 inclusive
- Combined two-party vote share as a proportion of all votes cast in election cycles from 2000 to 2012 inclusive
- For each age group of 20-45, 45-65, and 65+, the proportion of men and women in that age group who are:

DVN/42MVDX, https://doi.org/10.7910/DVN/42MVDX.

11. rrhelections.com, *Alaska Results by County Equivalent, 1960-2016 - RRH Elections*, December 17, 2020, accessed December 17, 2020, https://rrhelections.com/index.php/2018/02/02/alaska-results-by-county-equivalent-1960-2016/.

12. Alaska engages in a curious practice regarding its election results, one which is unique among the 50 states. Rather than report votes by county or county-equivalent, it reports by state legislative district. There are 40 state house districts and 30 county-equivalents, and their respective boundaries do not meaningfully overlap. Since the other data I had was at the county level, I needed to find a source for Alaska vote data at that geographical unit. rrhelections.com took the data for each election and constructed county level vote shares and totals for the major party candidates for all presidential elections from 1960 to the present. They presented this data as a series of .pngs of data tables, so I had to manually type the data into a spreadsheet to convert it to a usable form. However, considering that the alternative was to ignore the state of Alaska in my analysis, I feel justified in making this choice.

13. National Center for Family Marriage Research, *County-Level Marriage  Divorce Data, 2010*, December 16, 2020, accessed December 17, 2020, https://www.bgsu.edu/ncfmr/resources/data/original-data/county-level-marriage-divorce-data-2010.html.

14. County Health Rankings Project, *National Data  Documentation: 2010-2018 — County Health Rankings Roadmaps*, November 23, 2020, accessed December 17, 2020, https://www.countyhealthrankings.org/explore-health-rankings/rankings-data-documentation/national-data-documentation-2010-2018.

  - White
  - Black
  - Hispanic
  - Asian
  - American Indian or Alaskan Native
  - Native Hawaiian or Pacific Islander
  - Bi/multi-racial

- Total County Population
- Number of Marriages and Divorces (2010)
- Proportion of population with some college education
- Proportion of population not proficient in English
- Proportion of population who commute to work alone by car for more than 30 minutes daily
- Life expectancy quartile rank (within state)
- Quality of life quartile rank (within state)
- Quality of health behaviors quartile rank (within state) [15]
- Access to clinical care quartile rank (within state)
- Socioeconomic factors quartile rank (within state)[16]
- Median Household Income
- Physical environment quartile rank (within state)[17]

Broomfield County, Colorado (FIPS code 08014) was formed from parts of Adams, Boulder, Jefferson, and Weld counties in 2001. As such, it did not exist in the election data for the 2000 election. To solve this issue, I averaged the voting results for the four counties which formed Broomfield and inserted it into the data set at the appropriate index. Similarly, I replaced the Alaska observations included in the raw data set with county level data to maintain continuity across different data sources. For each county for a given year, I added the total votes earned by the Democratic candidate and the Republican candidate to find the total two-party vote, then calculated proportions of that vote for both parties.

The raw demographic data I used contained almost 750,000 rows (the unit of analysis for that raw data was county-year-age group). The size of this data set hindered my ability to process it, so I dropped all observations for years other than 2016 and recoded age from 18 categories to 3. To process age, I dropped all observations containing

15. i.e. prevalence of tobacco usage, adult obesity rates, rates of teenage pregnancy
16. i.e high school graduation rate, percentage of population in poverty, rates of violent crime
17. i.e. air quality, water quality, quality of housing

data for age groups unable to vote. Since one age group spanned 15-19 year olds, the youngest individuals reflected in my data set are 20 years old. I recoded the data into categories of 20-45, 45-65, and 65+, resulting in columns for the number of each racial category and gender for a given age (e.g. the number of Asian women aged 45-65). To standardize these numbers, I converted them to proportions of their age group (e.g. the proportion of 20-45 year olds who are white women). The choice of age categories was intended to reflect standard classifications roughly corresponding to young adults, middle-aged adults, and senior citizens.

# 5 Analysis

This analysis required the use of a number of machine learning tools found within the python package *scikit-learn* (often referred to as *sklearn*).[18] Once I was satisfied with the quality of my data set, I split it in two ways: first, into a single-variable outcome data set ("Y") and a multi-variable feature data set ("X"), and second, into a "training" set comprised of 75% of the observations in X and Y and a "test" set containing the remaining 25%. The split into training and test data requires the specification of a random state; since the data can be randomly divided an extremely large number of ways, specifying a random state means that the same division will be chosen every time, allowing results to be replicated.[19] This method allows machine learning models to be evaluated on their performance based on unfamiliar data, to gain a more complete understanding of the model's accuracy and predictive ability. After converting the data into training and test sets, the next step was to fill in any missing values with the mean of that variable. Doing this after the training-test split ensures that the test data is not in any way biased or affected by the training data.

The next step was to specify a k-fold generator. K-fold cross-validation is a machine learning technique which splits up a data set into a specified number of subsets, then processes each subset except for one (the "validation" subset) according to the specified algorithm. An error rate can be calculated for each tested subset, and then the average error rate becomes the cross-validation score. Like the training-test split, k-fold cross-validation also requires the specification of a random state to ensure that replication is possible.[20] Having created training and test data sets and specified a k-fold generator, the next step in this analysis was to scale all numerical variables using the preprocessing function *MinMaxScaler*. This function scales a numeric variable so that all values are between 0 and 1. It is especially useful in cases like this, where some variables reflect county populations which can vary significantly.

Finally, I specified the algorithms I wanted to test. For this analysis I used classifiers (machine learning tools which predict which one of a group of predetermined categories best describes a value of a discrete outcome variable) rather than regressors (which predict the value of a continuous outcome variable). I tested four algorithms for this analysis. The first is a naive bayes classifier. A naive bayes classifier calculates

18. F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research* 12 (2011): 2825–2830.

19. For this analysis, I specified random state = 149 for the training-test split.

20. For the k-fold cross-validation for this analysis, random state = 298.

the probability of a given observation belonging to a particular class based on the assumption that all feature variables are independent of each other. The second algorithm is a k-nearest-neighbors classifier. For a given observation a, this classifier calculates the $k$ observations closest to a and classifies a according to which class represents the plurality of the k observations.[21] The third algorithm is a decision tree classifier. Decision trees work by identifying splits in a feature variable which sort observations into branching categories. For example, a hypothetical decision tree based off of the data for this project might first divide the data based on whether unemployment levels in 2016 were above or below a certain threshold. Each of those divisions would then be divided further using other feature variables until the tree reaches a terminal state, typically specified by either a specific number of branches or a proportion of the data remaining in a split (splits in decision trees and related algorithms are called nodes). Finally, I tested a random forest classifier. Random forest algorithms operate under the assumption that different feature variables hold different levels of predictive power in a decision tree. As such, a random forest algorithm "grows" a specified number of decision trees, each one using a random set of feature variables (to ensure that highly predictive variables do not obscure the effects of other variables) and a specific number of nodes. These trees are then averaged to result in a final set of predictions.

The *sklearn* package includes a function called *GridSearch*, which allows for multiple algorithms with multiple parameter specifications to be run using a single command. *sklearn* also enables the creation of a machine learning pipeline, which incorporates the k-fold generator, the *MinMaxScaler* processing function, and the *GridSearch* algorithm tester in one function. Since this analysis is centered on classification, I chose to evaluate the accuracy of each algorithm using the ROC-AUC statistic, which compares the rate of true positives to false positives. I also independently calculated the number of misclassified counties relative to the size of the test data set, to aid with interpretability.

Furthermore, the *sklearn* package contains functions to facilitate the creation of partial dependency plots of the most important variables for each tested set of feature variables. These plots indicate the extent to which a machine learning algorithm depends on a particular feature variable to make accurate predictions. I also used the package *plotly* to create maps of the counties included in my test data, to visualize correct and incorrect predictions.[22]

# 6 Results

## 6.1 Specification 1: The Kitchen Sink[23]

The first machine learning model specification I tested included all 113 feature variables in my data set. Running the GridSearch returned a maximum AUC score of 0.9848, generated by the random forest algorithm. This score was obtained before I tuned any

21. i.e. For a given observation $a$ which can be "red" or "blue", if k equals 5, the classifier finds the 5 points closest to $a$. If three of those are red and two are blue, the classifier predicts that $a$ is red.

22. Plotly Technologies Inc., "Collaborative data science," 2015, https://plot.ly.

23. As in "everything but the kitchen sink."

of the parameters for any of the models. [24] Since the highest possible value for an AUC score is 1, I was curious as to which feature variables could be responsible for such a high score. To find this information, I tested the permutation importance of the feature variables (see Figure 1).[25]

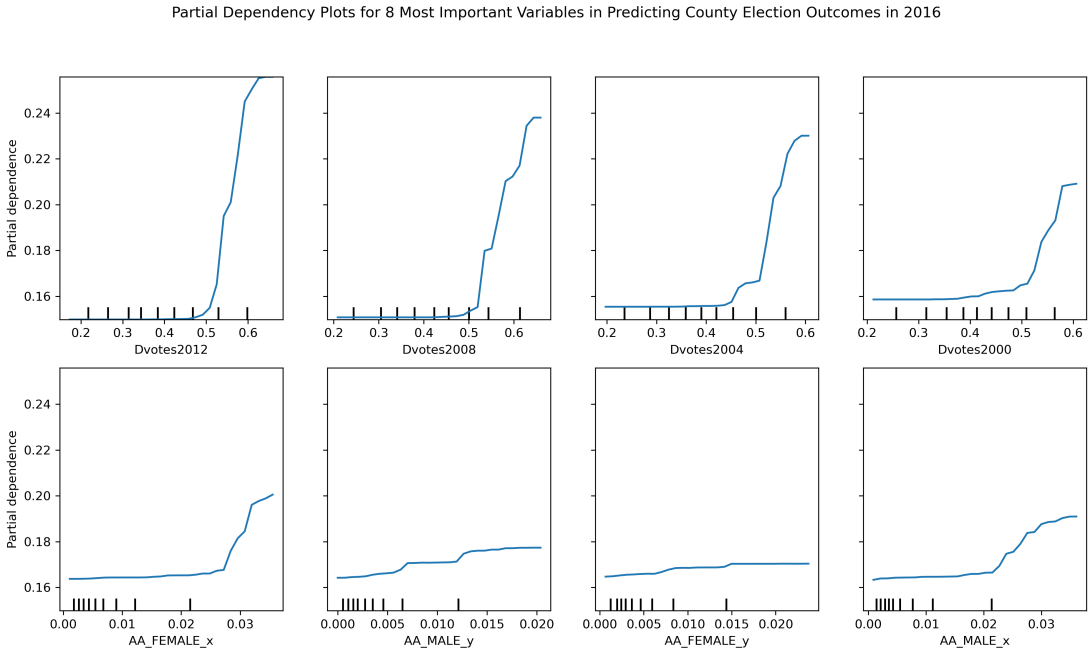


Figure 1: These figures depict the partial dependency plots for the eight most important variables in predicting which major party candidate won the plurality of votes in a county in 2016.

Figure 1 shows the partial dependency plots for the eight most important variables in this set of feature variables.[26] There is a clear visual difference in the vote share partial dependencies as compared to the demographic partial dependencies: the former are very steep, indicating that they exert a high degree of influence on the predicted probabilities (and thus potentially obscure other effects or relationships).

Figure 2 shows the counties or county-equivalents in the test data, color-coded by accuracy of prediction. 4.45% of the counties in the test data (35 out of 785) were misclassified; all 35 were predicted to be won by Trump. The dispersal of the misclassified counties is notable: of the 48 states represented in the training data, 19 had at least one

24. Ultimately, I chose not to tune any parameters for this project for two reasons: first, the initial accuracy achieved with a default set of parameters was over 95% for the first set of feature variables and over 92% for the second set, meaning that additional tuning efforts would almost certainly have diminishing marginal returns, and second, the overall goal of this project is to identify patterns in the errors, meaning that the quality of my results are not dependent on having a perfectly tuned model.

25. Roughly 50 of the feature variables were reported as having a permutation importance of 0, meaning that they did not affect the accuracy of the model.

26. The variables, in order: 1) Democratic two-party vote share in 2012, 2) Democratic two-party vote share in 2008, 3) Democratic two-party vote share in 2004, 4) Democratic two-party vote share in 2000, 5) the proportion of 20-45 year olds who are Asian women, 6) the proportion of 45-65 year olds who are Asian men, 7) the proportion of 45-65 year olds who are Asian women, and 8) the proportion of 20-45 year olds who are Asian men.

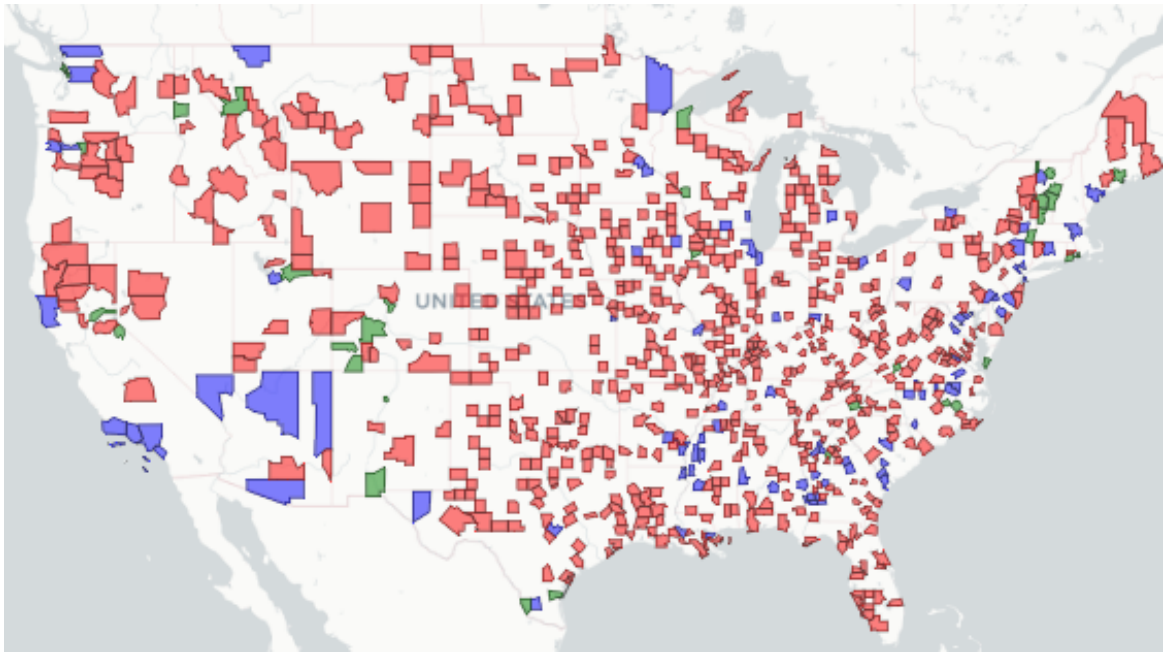

Figure 2: This map shows the counties included in the test data. Counties shaded red were correctly predicted as Trump victories, those shaded in blue were correctly predicted as Clinton victories, and those shaded in green were incorrectly predicted for Trump but were really won by Clinton. Due to space constraints, Alaska and Hawaii are not shown. Hawaii did not appear in the test data. Of the Alaska counties in the test data, only the borough of Sitka (FIPS code 02220), in the southernmost part of the Alaskan Penninsula, was misclassified for Trump instead of Clinton.

misclassified county. Only Vermont (6), Colorado (4), and North Carolina (3) had more than 2 errors.

Interestingly, the classifier did a particularly poor job at predicting the winning candidate in Vermont. Table 1 shows the predictions for the Vermont counties in the test data under a model specification which includes all feature variables. There are 14 counties in Vermont, 7 of which were sorted into the test data. 6 of those 7 were incorrectly predicted for Trump. The exception was Chittenden County (FIPS code 50007), the most populous county in the state (roughly one in four Vermont residents lives in Chittenden County).

Table 1: Predicted Outcomes for Selected Counties in Vermont

| FIPS | State | County | Winning Candidate in 2016 | Predicted Winner |
|---|---|---|---|---|
| 50003 | VT | Bennington County | Clinton | Trump |
| 50007 | VT | Chittenden County | Clinton | Clinton |
| 50013 | VT | Grand Isle County | Clinton | Trump |
| 50015 | VT | Lamoille County | Clinton | Trump |
| 50017 | VT | Orange County | Clinton | Trump |
| 50021 | VT | Rutland County | Clinton | Trump |
| 50027 | VT | Windsor County | Clinton | Trump |

## 6.2 Specification 2: Excluding Feature Variables Related to Voting

While it makes sense that past voting behavior would be more effective at predicting future outcomes than other factors,"partisanship in the United States is so static that the biggest predictor of which party will win a county is the party that won it four years ago" is a somewhat unsatisfying conclusion. As such, I decided to rerun the GridSearch, excluding the feature variables related to past voting outcomes. [27] The random forest algorithm again produced the highest AUC score of 0.9262 (as with the first iteration, this is without any significant attempt to tune the model). Figure 3 shows the partial dependency plots for the eight most important feature variables given this specification.[28]

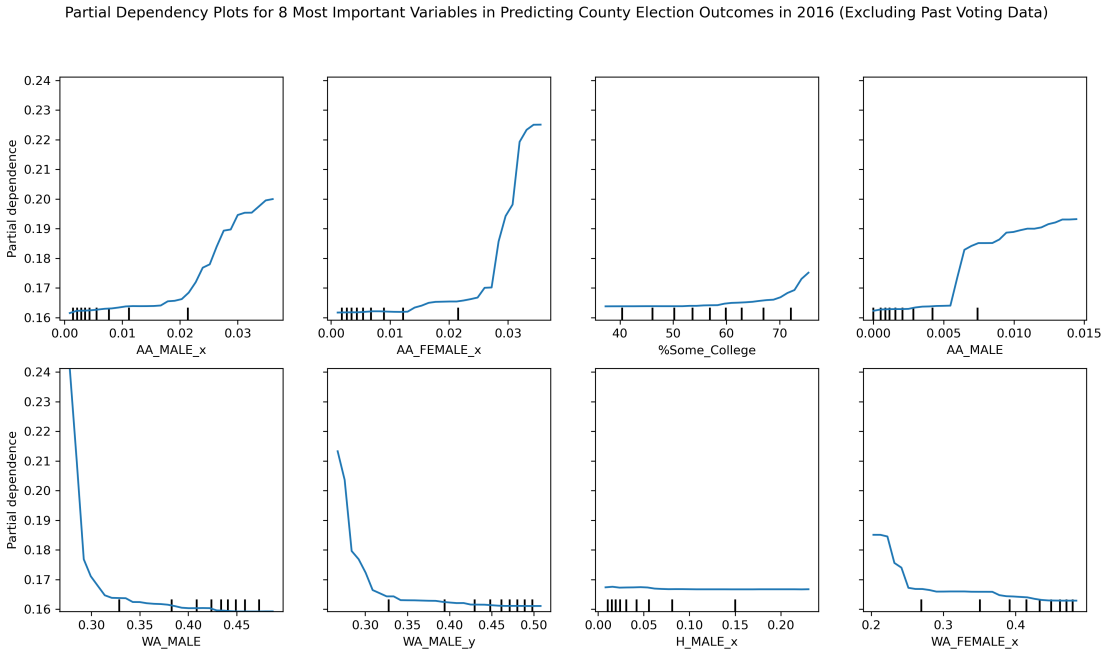


Figure 3: These figures depict the partial dependency plots for the eight most important variables in predicting which major party candidate won the plurality of votes in a county in 2016 for specification 2.

This model had an error rate of 7.26%, which represents an increase in error of 63% over the algorithm trained on the voting data from previous elections.[29] Of note are the relationships indicated by plots 5, 6, and 8, which all show the effects of various groups

27. These include the proportion of the two-party vote share won by the democrat in 2000, 2004, 2008,and 2012, the total number of votes cast in each of those elections, and the two-party vote total in those elections.

28. The eight variables depicted in Figure 3 are as follows: 1) the proportion of 20-45 year olds who are Asian men, 2) the proportion of 20-45 year olds who are Asian women, 3) the percentage of the adult population who has completed some college, 4) the proportion of adults older than 65 who are Asian men, 5) the proportion of adults older than 65 who are white men, 6) the proportion of 45-65 year olds who are white men, 7) the proportion of 20-45 year olds who are Hispanic men, and 8) the proportion of 20-45 year olds who are white women.

29. This makes sense, since this specification used the same random states as the "kitchen sink" but removed the most predictive variables, which would increase the error rate. In real terms, this means the algorithm misclassified 57 out of the 785 counties in the test data.

of white voters. Unlike other predictors shown in Figures 1 and 3, as the proportion of white voters in a county increases, that county becomes *less* likely to vote for the Democratic candidate.

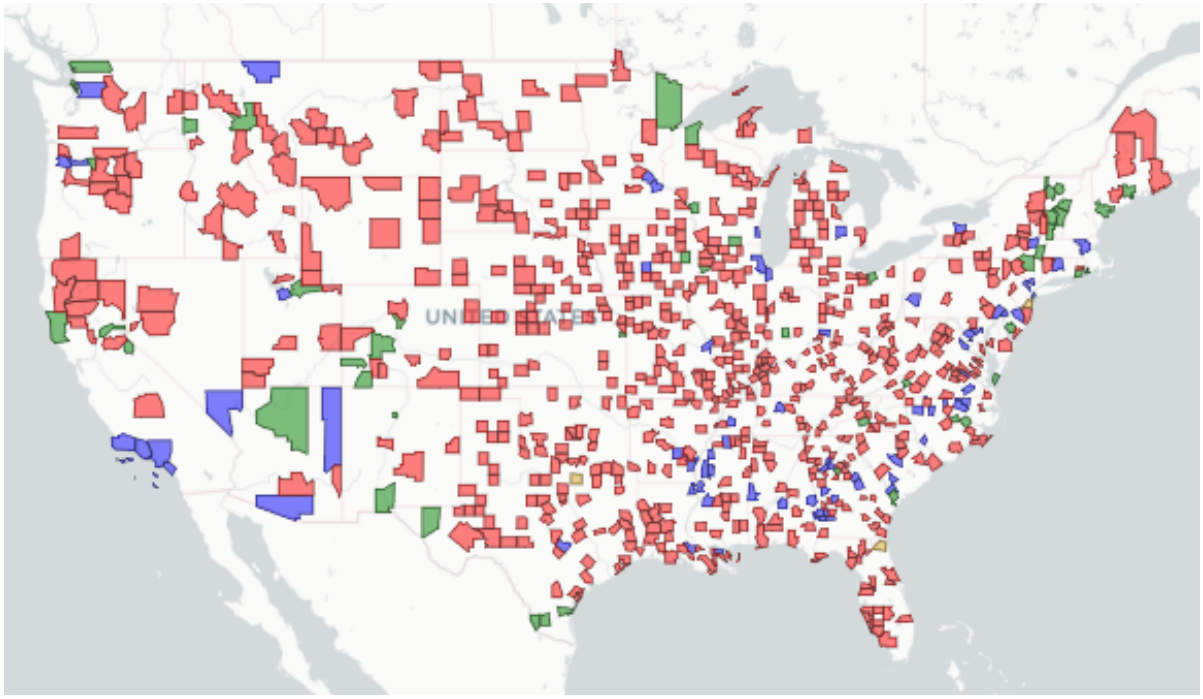

Figure 4: This map shows the counties included in the test data. Counties shaded red were correctly predicted for Trump, those shaded in blue were correctly predicted for Clinton, those shaded in green were incorrectly predicted for Trump but were really won by Clinton, and those shaded in yellow were incorrectly predicted for Clinton but were really won by Trump.

Figure 4 shows the counties included in the test data for this set of feature variables. As with the previous specification, the most successful algorithm trained on this data was much more likely to misclassify a county in favor of Trump than of Clinton, with 54 out of 57 erroneous classifications incorrectly predicting a Trump victory.[30]

I expect that this favorability toward Trump results from the real distribution of counties: Trump won three or four for every one won by Clinton. The modal outcome is a Trump victory, and so it is not wholly surprising that a machine learning algorithm would be more likely to incorrectly predict a Trump victory than incorrectly predict a Clinton victory. Overall, the number of misclassified counties is too small and geographically disparate to draw strong conclusions, but the particular case of Vermont merits discussion. The populations in counties which voted Republican in 2016 were roughly 10 points more white than those in Democratic-voting counties (regardless of age or gender). The demographic composition of Vermont more closely approximates the mean whiteness of Republican counties than Democratic ones. It is possible that this produced an effect which was sufficient to outweigh the predictive power of past voting patterns. The prevalence of Asian Americans as a predictor also deserves

30. The three counties in the test data which the algorithm incorrectly called for Clinton are Denton County, Texas, Duval County, Florida, and Monmouth County, New Jersey. (Since the random state was kept constant, the predictions for Alaska and Hawaii remain unchanged from those discussed in Figure 2.)

consideration, since that could be confounded by other variables. Exit polls from 2016 showed Clinton winning roughly 75% of Asian American voters,[31] compared to over 90% of Black voters. However, many of the states with the largest Black populations (Alabama, Mississippi, Georgia) are heavily Republican, while many of the states with the largest Asian populations (California, New York, New Jersey) are more Democratic. The overall effect of Asian voters could thus be concentrated relative to the effect of Black voters; further analysis is necessary to investigate this hypothesis.

# 7 Discussion

I reviewed my proposal to assess the success of this project, and have concluded that I did not achieve what I set out to do in my proposal. However, this requires clarification: the aim of my project (and the data I used) changed significantly between the time I wrote the proposal and the time I conducted this analysis and wrote this report. I initially planned to use data at the state-year level, but later switched to the county level (and focused on 2016 in particular) to increase the size of my data set. I ultimately ended up using data from a wider range of sources than I listed in my proposal. The decision to switch to county-level data allowed me to observe more granular trends than if I had used state-level data. My initial proposal also focused specifically on a link between partisan vote share (as the outcome variable) and unemployment rate, while the actual project focuses on which major-party candidate wins a given jurisdiction (and unemployment rate becomes one of a large number of feature variables, rather than the sole variable of interest). Perverse though it may sound, in considering this project I am glad to have "failed" in my original aims.

The primary way in which I would expand my analysis is to incorporate more data on education levels. I was able to find data on the proportion of a county population with some college education, but I was not able to find data on other levels of educational attainment at the county level without significant degrees of missingness. Additionally, I would have liked to more thoroughly investigate the misclassified counties for the two sets of feature variables used. While I was able to create maps of which counties were classified in particular ways, I did not have time to search for patterns in the erroneously classified counties beyond basic geography. Finally, given more time, I would like to test regressions on this data in addition to classifiers. Doing this would provide insight into how inaccurate these classifications were, and could potentially open up other lines of inquiry. In the aftermath of the election, much attention was given to three particular states considered to be pivotal in securing victory for Trump: Michigan, Pennsylvania, and Wisconsin. Figures 2 and 4 show that of the counties within these states which were included in the test data, only 2 or 3 counties (depending on the selected feature variables) were incorrectly classified; all of these were incorrectly predicted for Trump, and all are in Wisconsin. Future investigations focusing on predicting vote totals or vote shares could determine if this misclassification could significantly impact a prediction of this election.

31. "Clinton Won the Asian-American Vote, But Some Swing States Turned Toward Trump: Exit Poll," December 19, 2020, accessed December 19, 2020, https://www.nbcnews.com/news/asian-america/clinton-won-asian-american-vote-some-swing-states-turned-toward-n684716.

# References


Agriculture, U.S. Department of. *USDA ERS - Download Data*, October 30, 2020. Accessed October 30, 2020. https://www.ers.usda.gov/data-products/county-level-data-sets/download-data/.

Bureau, U.S. Census. *County Population by Characteristics: 2010-2019*, December 17, 2020. Accessed December 17, 2020. https://www.census.gov/data/tables/time-series/demo/popest/2010s-counties-detail.html.

"Clinton Won the Asian-American Vote, But Some Swing States Turned Toward Trump: Exit Poll," December 19, 2020. Accessed December 19, 2020. https://www.nbcnews.com/news/asian-america/clinton-won-asian-american-vote-some-swing-states-turned-toward-n684716.

Data, MIT Election, and Science Lab. *U.S. President 1976–2016*. V. V5, 2017. https://doi.org/10.7910/DVN/42MVDX. https://doi.org/10.7910/DVN/42MVDX.

Family Marriage Research, National Center for. *County-Level Marriage  Divorce Data, 2010*, December 16, 2020. Accessed December 17, 2020. https://www.bgsu.edu/ncfmr/resources/data/original-data/county-level-marriage-divorce-data-2010.html.

Fiorina, Morris P. "The Political Parties Have Sorted." In *Essays on Contemporary American Politics*, 1–20. Stanford CA: Hoover Institution Press, 2016.

Frey, William H. "Census shows pervasive decline in 2016 minority voter turnout," December 18, 2020. Accessed December 18, 2020. https://www.brookings.edu/blog/the-avenue/2017/05/18/census-shows-pervasive-decline-in-2016-minority-voter-turnout/.

Inc., Plotly Technologies. "Collaborative data science," 2015. https://plot.ly.

Kennedy, Courtney, Mark Blumenthal, Scott Clement, Joshua D Clinton, Claire Durand, Charles Franklin, Kyley McGeeney, et al. "An Evaluation of the 2016 Election Polls in the United States." *Public Opinion Quarterly* 82, no. 1 (February 2018): 1–33. ISSN: 0033-362X. https://doi.org/10.1093/poq/nfx047. eprint: https://academic.oup.com/poq/article-pdf/82/1/1/24265180/nfx047.pdf. https://doi.org/10.1093/poq/nfx047.

McKee, Seth C. "Rural Voters and the Polarization of American Presidential Elections." *PS: Political Science and Politics* 41, no. 1 (2008): 101–108. ISSN: 10490965, 15375935. http://www.jstor.org/stable/20452117.

Park, Taeyong, and Andrew Reeves. "Local Unemployment and Voting for President: Uncovering Causal Mechanisms," September 28, 2018. Accessed December 18, 2020. https://andrewreeves.org/papers/unemployment.pdf.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, et al. "Scikit-learn: Machine Learning in Python." *Journal of Machine Learning Research* 12 (2011): 2825–2830.

Project, County Health Rankings. *National Data  Documentation: 2010-2018 — County Health Rankings  Roadmaps*, November 23, 2020. Accessed December 17, 2020. https://www.countyhealthrankings.org/explore-health-rankings/rankings-data-documentation/national-data-documentation-2010-2018.

rrhelections.com. *Alaska Results by County Equivalent, 1960-2016 - RRH Elections*, December 17, 2020. Accessed December 17, 2020. https://rrhelections.com/index.php/2018/02/02/alaska-results-by-county-equivalent-1960-2016/.