Using machine learning to predict county preferences for U.S. Presidential elections

Steven A. Jordan
DSC 540 | DePaul University

Abstract

Every year, the US Census Bureau conducts detailed demographic, economic, and business data surveys in every county in the United States. Political organizations and candidates can (and do) use this information in order to build comprehensive strategies to better understand voters, serve their people, and focus their efforts while campaigning in elections. The goal of this project is, by using data obtained by the US Census Bureau, to build a model that will predict the margin between the percentage of Democratic and Republican votes, and to identify its corresponding most important features. Ideally, this model should be useful for anyone working in politics to provide insight as to where to focus resources. Multiple supervised machine learning algorithms (LASSO, Ridge, Random Forest, Gradient Boosting, and Support Vector Machine regression) were conducted, with the Gradient Boosting-derived model to have the best overall performance. However, the error rate is still high enough that the utility of this model may be minimal, and the model should be further developed.

Key Words: regression, LASSO, Ridge, Random Forest, Gradient Boosting, Support Vector Machines, elections, US presidential elections, US Census, demographic analysis

1. Introduction

1.1 Problem

Ever increasingly, data analysis is becoming a mainstream part of our political landscape. From analytical journalism (Solop, 2016), to reapportioning districts using gerrymandering (Thomas, 1989), to targeting political ads on Facebook (Tufekci, 2014) – predictive models are being used everywhere. And next year, as a result of the 2020 census, the United States will obtain a tremendous amount of new data regarding the shifts in population, age, race, and other demographics on a detailed county-by-county level (US Census Bureau, 2020 Census Questionnaire, 2019). Furthermore, the US Census Bureau updates demographic data annually via the Population Estimates Program, and through surveys like the American Community Survey (ACS), the Survey of Business Owners and Self-Employed Persons (SBO), and the County Business Patterns (CBP) surveys (US Census Bureau, Index A-Z, 2019). While the full 2020 Census results are not scheduled to be released until December 2020, after the Presidential election, this data will certainly be used by the Republican and Democratic National Committees, individual political campaigns, and think-tanks to determine how to focus their resources and efforts over the next decade.

1.2 Goal

In this project, I attempted to build models that predict the percentage difference between the Democratic and Republican vote share of any county by using data from the United States Census Bureau, and the results from the 2012 and 2016 US Presidential Elections. These models are regression-based (not classification-based), however, the results can very easily be classified into a win for the Republican or Democratic party. I will not account for the percentage of votes for any third parties (e.g. Libertarian, Green, etc.) as they only averaged a combined 3.7% of the popular vote in the last two elections, without winning a single county (Azhar, 2016). If successful, this model could be used by any political party in preparation for not just the 2020 (using 2019 estimates) Presidential Election but also can use the most important features to develop a strategy to build in-roads to critical communities over the next decade.

1.3 Related Work

While there is plenty of published work that uses general polling or social media sentiment analysis to predict elections, there is surprisingly very little related work that uses US Census or demographic data to build predictive models – or at least models that are more complex than logistic or ordinary-least-squares (OLS) regression analyses. Two studies (Flaxman et al., 2016 and Pesta & Mcdaniel, 2014) that conducted OLS

regression on state-level demographic data found that the white population size and education attainment were two of the three the most important factors in predicting the vote share between the Republican and Democratic candidates in elections since 2000 (with the latter also finding very high importance to religiosity). Similarly, the study *Blue City … Red City? A Comparison of Competing Theories of Core County Outcomes in U.S. Presidential Elections, 2000-2012* (Ambrosius, 2016) analyzes how different demographics of urban counties are more or less likely to influence the partisan swing of the county in an election. Their analysis produces some of the same most important features my machine learning models produced (e.g. % of population that is black, population size). Two features that resulted with very high regression coefficients in this study that were not included in my own analysis were: % of same-sex households, and % evangelicals.

The paper *US Presidential Election 2012 Prediction using Census Corrected Twitter Model* (Choy et al., 2012) constructed a model using sentiment analysis of tweets leading up to the 2012 election; then using data regarding household internet access (provided by the US Census Bureau), was able to weight their predictions and increase accuracy. This leads me to believe, it could be possible to incorporate a social media sentiment analysis in combination with a demographic-based model for a more robust forecasting tool. The paper *Advertising Effects in Presidential Elections* (Gordon & Hartmann, 2013) uses OLS regression to determine the impact of advertising in a US Presidential election, and concludes that had there been no mainstream advertising in the 2000 election, three more states would have voted for Gore, flipping the result of the election. In conjunction with my research, a political organization like the Democratic or Republican National Committee, could advertise within counties in which they are under-performing (according to the model) – to potentially change an election outcome.

2. Methodology

2.1 Data Sourcing

The county-by-county election result data for the 2012 and 2016 US Presidential elections were sourced via a Kaggle repository (Wilson, 2018) - which in turn sourced the data from the online journal *The Guardian*. County sizes were also sourced from the Kaggle repository. The dependent variables for each county in each data set is: **% of Democratic votes - % of Republican votes** (thus, a positive value is a Democratic win, and a negative value is a Republican win). The county demographic and business features were sourced via searching the data portal on the US Census Bureau website (Data Access and Dissemination Systems, 2019); I downloaded 36 data sets with county-level demographic data for the years preceding an election (2011 and 2015) and business data for the year 2012 (prior to the election) from the 2010 Census, the 2011 and 2015 ACS, the 2011 and 2015 SBO, and the 2012 CBP survey.

Using the US Census QuickFacts page as a guide (US Census Bureau, QuickFacts, 2019), I created two subsets of data (some features on the QuickFacts page were omitted due to a high volume of missing values): a 2012 election data set, a 2016 election data set. The 2012 data set uses the 2011 ACS and SBO survey values; the 2016 data set uses 2015 ACS and SBO survey values; and both use the 2012 CBP survey data. Each data set has forty-six features: twenty-two of which are interval-numeric, twenty-four of which are ratio-numeric. In general, the features can be categorized into seven areas: Population, Age/Sex, Race/Ethnicity, Housing/Living Arrangements, Education, Economy/Business, and Land Area. A full dictionary of the features can be found in **Appendix 1**.

Because Alaska has only one county used for voter counts, but multiple boroughs used by the US Census, the state's data did not line up, and all locations in Alaska were removed. Also, because Shannon County, SD was created in 2015, it was not included. These removals results in 3111 remaining counties for analysis. Seven of the features had less than 2% missing values, one feature had about 7% missing values, and one feature had 18% missing values. Each of these features (described in **Appendix 1)** had their missing values replaced by that feature's mean.

2.2 Data Transformation and Feature Selection

The distribution of the dependent variable (difference in vote percentage) had a nearly normal distribution for both 2012 and 2016 – as seen in **Figure 1**. There is a slight right skew in both distributions,

indicating a greater proportion of GOP-winning counties, but that was not unexpected. However, many of the demographic features had extreme skews, as seen in **Figure 2**; for this reason the features were log-transformed, resulting in a more normal distribution. The features were also scaled using min-max scaling to reduce bias towards the heavier ratio-type features.

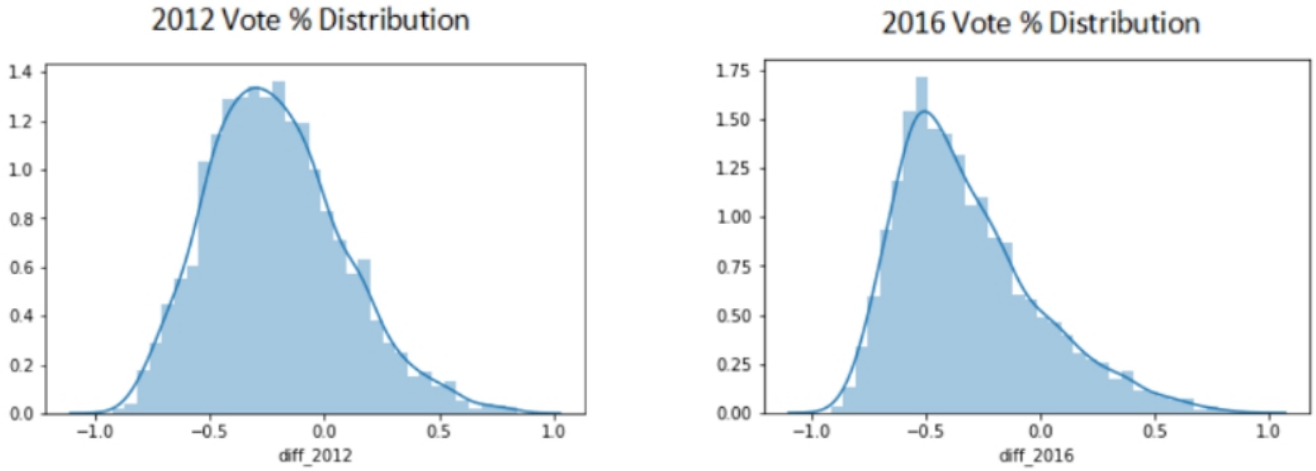


Figure 1. Distribution of the vote percentage difference in the 2012 and 2016 US Presidential elections.

Figure 2. Distribution of the number of female-owned firms, before (left) and after (right) log-transformation. This distribution was very typical among the ratio-type attributes.

After transforming the data, the correlation between the features was analyzed. If one looks at 2012 correlation plot in **Figure 3**, one can see that there are groups of features which have nearly 100% correlation with one another. The remaining correlation plots can be seen in **Appendix 2**. Five features, which had nearly perfect correlation with another predictive feature, were removed – resulting in forty-one remaining variables. The removed features, and their highly-correlated remaining feature can be seen in **Table 1.** Several highly-correlated variables remained in the data sets since there will be another round of wrapper-based feature selection when model-building.

The records of two data sets were randomly shuffled and split into 80% training and 20% test subsets. The two training sets and two test sets were also combined with one another, forming a third subset: a combined 2012/2016 election training and test set. I wanted to create these three sets in order to test the generalization of any models built; after all, if a model performs well in the 2012 test set, but performs poorly on the 2016 or combined test sets, it is an indication that the model is overfit to that election year and not generalizable.

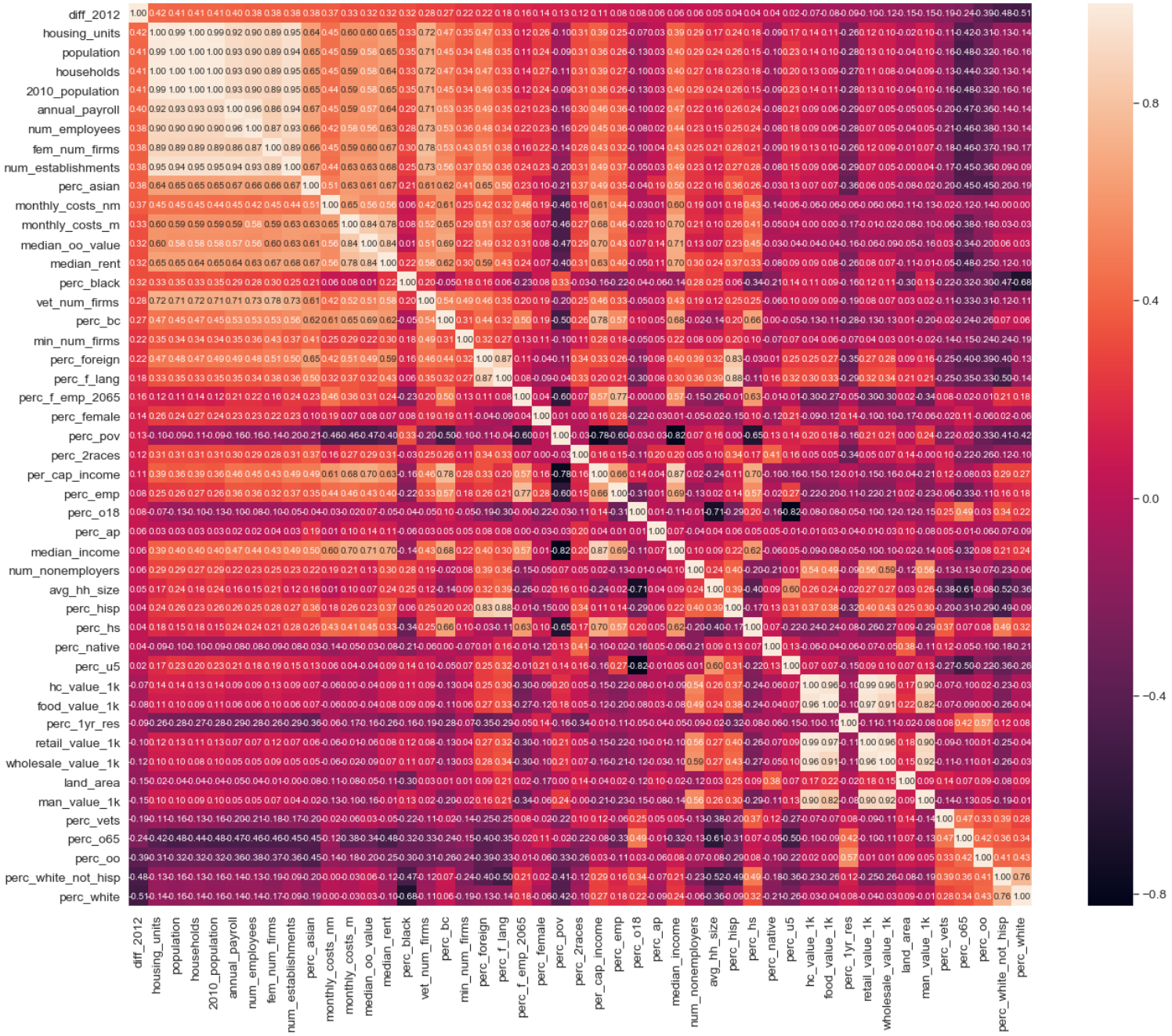


Figure 3. Heatmap and correlation plot of the 2012 data set, after log-transformation. Notice the groupings of nearly perfectly correlated variables.

Removed Feature	Remaining Nearly Perfectly Correlated Feature
# Housing Units	(2011 or 2015) Population Estimate
# Households	(2011 or 2015) Population Estimate
2010 Census Population	(2011 or 2015) Population Estimate
# of Employees	Annual Payroll
Total retail sales	Total health care and social assistance revenue

Table 1. List of the manually removed features (due to nearly perfect correlation) and their corresponding feature that remained in the data sets.

2.3 Machine Learning Approaches and Evaluation

To create the best possible model with the demographic data available, I constructed models using five different supervised learning algorithms and compared the results of the test data. The five algorithms are described in brief detail in **Table 2**. To evaluate the models, the mean average error (MAE), the root mean squared error (RMSE), and the explained variance were calculated and compared.

Algorithm	Algorithm Description	Python Package Used
LASSO Regression (LR)	Linear regression technique that handles multicollinearity by using a shrinkage estimator and using L1 regularization, which reduces all coefficients by the absolute value of the magnitude of the coefficients. The regularization penalty can be tuned to fit the best performing model. (Melkumova & Shatskikh, 2017)	LassoCV from scikit-learn. Cross validation is incorporated.
Ridge Regression (RR)	Linear regression technique that handles multicollinearity by using a shrinkage estimator and using L2 regularization, which reduces all coefficients by the square of the magnitude of the coefficients. The regularization penalty can be tuned to fit the best performing model. (Melkumova & Shatskikh, 2017)	RidgeCV from scikit-learn. Cross validation is incorporated.
Gradient Boosting (GB)	An ensemble approach that trains numerous regression trees in succession. After each tree's accuracy is evaluated, it determines which observations were the most difficult to predict, and which were the easiest. For each successive tree, the weight of the difficult-to-predict observations is increased, and the weight of the easy-to-predict observations is decreased. The final prediction is calculated using the weighted average of the different individual trees. (Ogutu et al., 2011)	GradientBoostingRegressor from scikit-learn. Cross-validation is performed separately.
Random Forest (RF)	An ensemble of individual regression trees, each of which producing their own prediction (Yiu, 2019). The trees' predictions are averaged, and the average becomes the overall model's prediction. Random forests use a combination of bootstrap aggregation and random feature selection at tree node splits to reduce the correlation of its individual trees, thereby increasing the generalization of the overall model. (Ogutu et al., 2011)	RandomForestRegressor from scikit-learn. Cross-validation is performed separately.
Support Vector Machine Regression (SVM)	Support vector machines arrays predictors in an observational space using a set of inner products, transforming the data into the required form via a kernel (Ogutu et al., 2011)	SVR from scikit-learn. Cross-validation is performed separately.

Table 2 – Description of the algorithms used in model creation.

Prior to model construction, I utilized each algorithms' scikit-learn wrapper-based feature selection method to reduce the number of features for analysis. The features selected for each algorithm can be seen in **Appendix 3**. For the SVM, RF, and GB approaches, a wide parameter grid was created, and I followed a procedure that first implemented the RandomizedSearchCV function followed by the GridSearchCV function from scikit-learn to tune the parameters and produce the best possible scores. For LASSO and Ridge Regression, the penalty parameters were manually tuned.

3. Results and Discussion

After training and tuning models for each data set with every algorithm, each model was tested on each of the test sets. Multiple test sets were used in order to better understand the generalizability of the models. **Table 3** shows a comparison of scores from the different fifteen different models created.

		2012 Test Set			2016 Test Set			Combined Test Set		
		RMSE	MAE	Exp Var	RMSE	MAE	Exp Var	RMSE	MAE	Exp Var
Lasso Regression	2012	0.22	0.15	0.42	1.03	1.00	0.15	0.75	0.58	0.00
	2016	0.27	0.22	0.55	0.16	0.12	0.73	0.22	0.17	0.56
	Combined	**0.18**	**0.14**	**0.60**	**0.16**	**0.12**	**0.73**	**0.17**	**0.14**	**0.67**
Ridge Regression	2012	0.21	0.16	0.48	0.83	0.80	0.34	0.61	0.48	0.00
	2016	0.25	0.21	0.56	0.16	0.12	0.73	0.21	0.17	0.58
	Combined	**0.19**	**0.15**	**0.59**	**0.15**	**0.12**	**0.73**	**0.17**	**0.14**	**0.67**
Gradient Boosting	2012	**0.14**	**0.10**	**0.78**	**0.16**	**0.12**	**0.73**	**0.15**	**0.11**	**0.76**
	2016	0.20	0.16	0.53	0.15	0.11	0.76	0.18	0.14	0.66
	Combined	0.18	0.14	0.62	0.15	0.12	0.74	0.17	0.13	0.70
Random Forest	2012	**0.14**	**0.11**	**0.76**	**0.17**	**0.14**	**0.70**	**0.16**	**0.13**	**0.73**
	2016	0.20	0.16	0.52	0.16	0.12	0.74	0.18	0.14	0.64
	Combined	0.18	0.14	0.62	0.15	0.12	0.75	0.17	0.13	0.70
Support Vector Regression	2012	0.15	0.12	0.74	0.68	0.63	0.00	0.50	0.37	0.00
	2016	0.27	0.23	0.45	0.12	0.09	0.83	0.21	0.16	0.58
	Combined	**0.15**	**0.11**	**0.75**	**0.12**	**0.09**	**0.84**	**0.13**	**0.10**	**0.80**

Table 3. A comparison of the scores generated from the different models trained and tested on the 2012, 2016, and combined data sets. The best performing model of each algorithm is highlighted in green.

The two tree-based ensemble methods (GB and RF) performed very similarly well, with very similar scores for all the test sets. However, the Gradient Boosting models performed slightly better than Random Forest models overall. Very surprisingly, the RF and GB models trained on only the 2012 data sets outperformed the RF and GB models trained on the combined data set. Because all models have similar scores across the different test sets, it leads me to believe that the ensemble models are very generalizable to election years yet untested.

The two linear regression methods and the support vector regression had the poorest scores for models trained on *only* the 2012 or 2016 data, but, unsurprisingly, the models trained on the combined data set performed the much better. In fact, the SVM model trained on the combined data set had the overall best scores, even when compared to the ensemble models. But, because the 2012 and 2016 SVM models performed so poorly on data sets on which they were not trained, I believe there is a strong likelihood that the SVM combined model is overfit, and would not be generalizable to other election years. It is for this reason that I believe the best model is the Gradient Boosting trained on the 2012 data. While its scores are not as high as the combined SVM model, all three GB models perform consistently across the different test sets on which they were not trained. This is a very strong indicator that the GB models are the most generalizable.

The parameters and the most important features for each algorithm's best performing model can be viewed in **Appendix 4**. The five most important features for 2012 GB model are the % of black residents, % of white residents, the monthly cost of living for people with a mortgage, the overall population size, and the % of residents over 18 years old. This model-building process is a great example that the most correlated features to the dependent variable are not necessarily the most important predictors; **Table 4** shows the ranking of these five features in raw correlation to the dependent variables (% difference in 2012 and 2016).

	Best GB Model	% Diff in 2012	% Diff in 2016
% Black	1	12	14
% White	2	1	1
Monthly Costs (People w/ mortgage)	3	10	12
Population	4	3	6
% over 18	5	31	36

Table 4. The ranking of the five most important features in the best GB model in the correlation to the dependent variables of the 2012 and 2016 data sets.

Interestingly enough, the most important features for the model are very different than the most correlated features relative to the dependent variables. The only feature that is in the top five for all three rankings is the % of white residents (although overall population is close). And shockingly, the feature representing the % of residents over 18 (or stated differently – a lower % of children), was one of the least correlated features to the dependent variables, but was the fifth most important feature to the model.

4. Conclusion

4.1 Summary

While I am confident that I constructed one of the best possible models for the data set used, I am not confident that this is the best possible model to achieve the goal of predicting the margin between the percentage of Democratic voters and Republican voters. The lowest mean average error was 0.11, or a 11% difference between the votes. Seeing how that error can go in either direction, that is a very wide berth for mistakes when a fraction of a point decides the winner. However, I believe the most important features I found can still provide some insight to people working on campaigns – whether it's for advertising or for direct voter outreach.

4.2 Future Work

As per future work, I have several recommendations on how to improve the model. If possible, one should also incorporate demographic data on religions and overall religiosity, and for same-sex households. They were not part of my data sets (since they were not listed on the the US Census' QuickFacts page), but other studies listed in the **Related Work** section indicate that they could be very predictive features. Another feature that was omitted, because of a high count of missing values before 2015, was broadband internet access; this feature could be important in future models when the data is more consistently collected. Ideally, one would also test the models on the 2020 US Presidential election, and pre-2012 elections. Furthermore, I recommend exploring more deeply as to why the Gradient Boosting and Random Forest models trained on only the 2012 data sets outperformed the models trained on both the 2012 and 2016 training sets combined. One should investigate if there is a specific feature that was selected, or if there is an extreme observation within the 2016 data set that is throwing the algorithms off.

References

1. Ambrosius, J. D. (2016). Blue City … Red City? A Comparison of Competing Theories of Core County Outcomes in U.S. Presidential Elections, 2000–2012. *Journal of Urban Affairs*, *38*(2), 169–195. doi: 10.1111/juaf.12184
2. Azhar, H. (2016, December 29). 2016 Vs. 2012: How Trump's Win And Clinton's Votes Stack Up To Romney And Obama. Retrieved from https://www.forbes.com/sites/realspin/2016/12/29/2016-vs-2012-how-trumps-win-and-clintons-votes-stack-up-to-obama-and-romney/#46e54e8b1661.
3. Choy, M., Cheong, M., Laik, M. N., & Shung, K. P. (2012). US Presidential Election 2012 Prediction using Census Corrected Twitter Model. *Eprint ArXiv:1211.0938*. Retrieved from https://arxiv.org/abs/1211.0938
4. Data Access and Dissemination Systems (DADS). (2019). American FactFinder. Retrieved October 10, 2019, from https://factfinder.census.gov/faces/nav/jsf/pages/index.xhtml.
5. Flaxman, S., Sutherland, D. J., Wang, Y.-X., & Teh, Y. W. (2016). Understanding the 2016 US Presidential Election using ecological inference and distribution regression with census microdata. *ArXiv:1611.03787 [Stat.AP]*.
6. Gordon, B. R., & Hartmann, W. R. (2013). Advertising Effects in Presidential Elections. *Marketing Science*, *32*(1), 19–35. doi: 10.1287/mksc.1120.0745
7. Melkumova, L., & Shatskikh, S. (2017). Comparing Ridge and LASSO estimators for data analysis. *Procedia Engineering*, *201*, 746–755. doi: 10.1016/j.proeng.2017.09.615

8. Ogutu, J. O., Piepho, H.-P., & Schulz-Streeck, T. (2011). A comparison of random forests, boosting and support vector machines for genomic selection. *BMC Proceedings*, *5*(S3). doi: 10.1186/1753-6561-5-s3-s11
9. Pesta, B. J., & Mcdaniel, M. A. (2014). State IQ, well-being and racial composition as predictors of U.S. presidential election outcomes. *Intelligence*, *42*, 107–114. doi: 10.1016/j.intell.2013.11.006
10. Solop, F. I., & Wonders, N. A. (2016). Data Journalism Versus Traditional Journalism in Election Reporting. *Electronic News*, *10*(4), 203–223. doi: 10.1177/1931243116656717
11. Thomas, S. J. (1989). The Lack of Judicial Direction in Political Gerrymandering: An Invitation to Chaos Following the 1990 Census. *Hastings Law Journal*, *40*(5), 1067–1093. Retrieved from https://repository.uchastings.edu/cgi/viewcontent.cgi?article=2981&context=hastings_law_journal
12. Tufekci, Z. (2014). Engineering the public: Big data, surveillance and computational politics. *First Monday*, *19*(7). doi: 10.5210/fm.v19i7.4901
13. US Census Bureau. (2019, October 28). 2020 Census Questionnaire. Retrieved from https://www.census.gov/programs-surveys/decennial-census/technical-documentation/questionnaires/2020.html.
14. US Census Bureau. (2019, August 15). Index A-Z. Retrieved from https://www.census.gov/about/index.html.
15. US Census Bureau. (2019). QuickFacts: United States. Retrieved from https://www.census.gov/quickfacts/fact/table/US/PST045218.
16. Wilson, J. (2018, September 10). 2012 and 2016 Presidential Elections. Retrieved October 1, 2019, from https://www.kaggle.com/joelwilson/2012-2016-presidential-elections.

Appendix

Appendix 1. Data dictionary

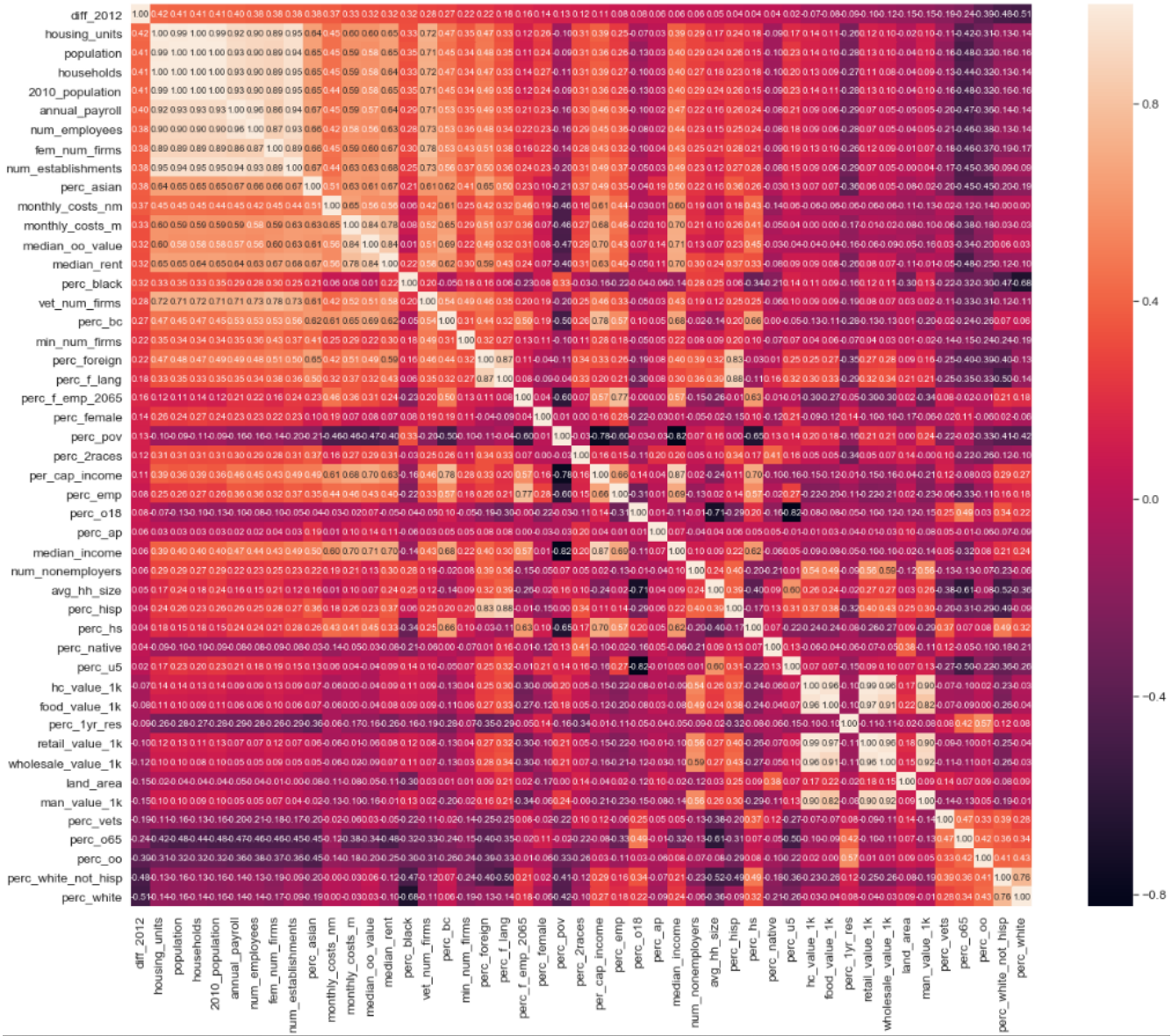
Variable Name	Type	Definition	Source	Missing Values
fips	index	Unique code assigned to every county. Used as dataframe index for 2012, 2016 election dataframes; and the 2012 election rows of the combined dataframe	2010 US Census	0
county_name	index	Name of each county. Used as the index of the 2016 election rows of the combined dataframe	2010 US Census	0
diff_2012	dependent variable	The percentage of a county's votes that went to the Republican candidate subtracted from the percentage that went to the Democratic candidate, 2012	Kaggle	0
diff_2016	dependent variable	The percentage of a county's votes that went to the Republican candidate subtracted from the percentage that went to the Democratic candidate, 2016	Kaggle	0
2010_population	ratio	Population, Census, April 1, 2010	2010 US Census	0
land_area	ratio	Land area in square miles, 2010	Kaggle	0
population	ratio	Population estimates, 2011 or 2015	American Community Survey (2011 or 2015)	0
perc_female	interval	Female persons, percent	American Community Survey (2011 or 2015)	0
perc_u5	interval	Persons under 5 years, percent	American Community Survey (2011 or 2015)	0
perc_o18	interval	Persons over 18 years, percent	American Community Survey (2011 or 2015)	0
perc_o65	interval	Persons 65 years and over, percent	American Community Survey (2011 or 2015)	0
perc_2races	interval	Two or More Races, percent	American Community Survey (2011 or 2015)	0
perc_white	interval	White alone, percent	American Community Survey (2011 or 2015)	0

Variable Name	Type	Definition	Source	Missing Values
perc_black	interval	Black or African American alone, percent	American Community Survey (2011 or 2015)	0
perc_native	interval	American Indian and Alaska Native alone, percent	American Community Survey (2011 or 2015)	0
perc_asian	interval	Asian alone, percent	American Community Survey (2011 or 2015)	0
perc_ap	interval	Native Hawaiian and Other Pacific Islander alone, percent	American Community Survey (2011 or 2015)	0
perc_hisp	interval	Hispanic or Latino, percent	American Community Survey (2011 or 2015)	0
perc_white_not_hisp	interval	White alone, not Hispanic or Latino, percent	American Community Survey (2011 or 2015)	0
perc_vets	interval	Veterans, percent	American Community Survey (2011 or 2015)	0
perc_foreign	interval	Foreign born persons, percent	American Community Survey (2011 or 2015)	0
housing_units	ratio	Housing units	American Community Survey (2011 or 2015)	0
perc_oo	interval	Owner-occupied housing units, percent	American Community Survey (2011 or 2015)	0
median_oo_value	ratio	Median value of owner-occupied housing units	American Community Survey (2011 or 2015)	0
monthly_costs_m	ratio	Median selected monthly owner costs -with a mortgage	American Community Survey (2011 or 2015)	0
monthly_costs_nm	ratio	Median selected monthly owner costs -without a mortgage	American Community Survey (2011 or 2015)	0
median_rent	ratio	Median gross rent	American Community Survey (2011 or 2015)	0
households	ratio	Households	American Community Survey (2011 or 2015)	0
avg_hh_size	ratio	Persons per household	American Community Survey (2011 or 2015)	0
perc_1yr_res	interval	Living in same house 1 year ago, percent	American Community Survey (2011 or 2015)	0
perc_f_lang	interval	Language other than English spoken at home, percent	American Community Survey (2011 or 2015)	0
perc_hs	interval	High school graduate or higher, percent of persons age 25 years+	American Community Survey (2011 or 2015)	0
perc_bc	interval	Bachelor's degree or higher, percent of persons age 25 years+	American Community Survey (2011 or 2015)	0
perc_emp	interval	In civilian labor force, total, percent of population age 16 years+	American Community Survey (2011 or 2015)	0
perc_f_emp_2065	interval	In civilian labor force, female, percent of population ages 20-65	American Community Survey (2011 or 2015)	0
num_nonemployers	ratio	Total nonemployer establishments	County Business Patterns (2012)	0
man_value_1k	ratio	Total manufacturers shipments ($1,000)	County Business Patterns (2012)	1
wholesale_value_1k	ratio	Total merchant wholesaler sales ($1,000)	County Business Patterns (2012)	0
retail_value_1k	ratio	Total retail sales ($1,000)	County Business Patterns (2012)	0
hc_value_1k	ratio	Total health care and social assistance receipts/revenue ($1,000)	County Business Patterns (2012)	0
food_value_1k	ratio	Total accommodation and food services sales ($1,000)	County Business Patterns (2012)	0
median_income	ratio	Median household income	American Community Survey (2011 or 2015)	4
per_cap_income	ratio	Per capita income in past 12 months	American Community Survey (2011 or 2015)	4

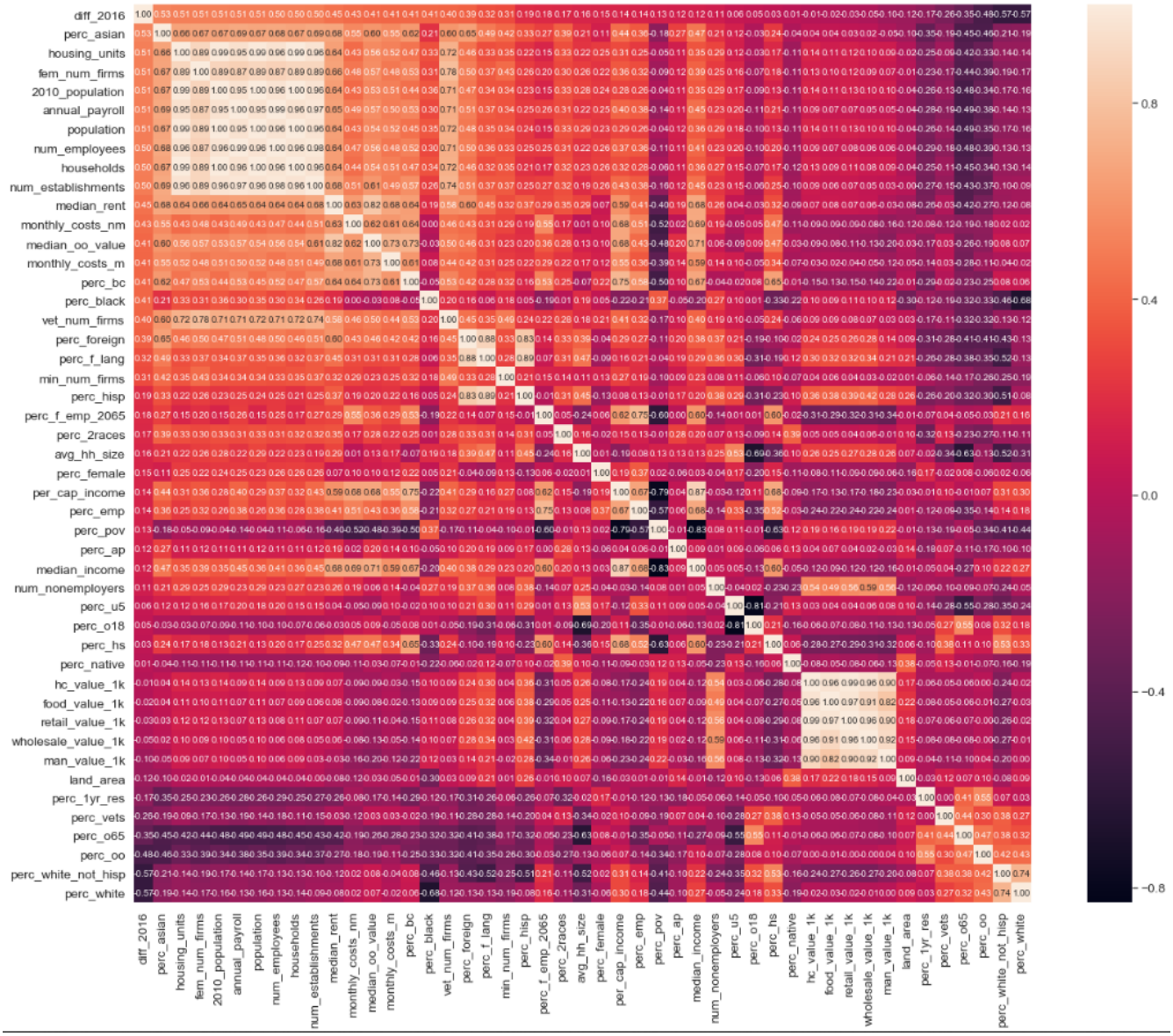
Variable Name	Type	Definition	Source	Missing Values
perc_pov	interval	Persons in poverty, percent	American Community Survey (2011 or 2015)	4
num_establishments	ratio	Total employer establishments	Survey of Business Owners and Self-Employed Persons (2011 or 2015)	0
num_employees	ratio	Total employment	Survey of Business Owners and Self-Employed Persons (2011 or 2015)	52 (2011), 2 (2015)
annual_payroll	ratio	Total annual payroll ($1,000)	Survey of Business Owners and Self-Employed Persons (2011 or 2015)	30 (2011), 2 (2015)
vet_num_firms	ratio	Veteran-owned firms	Survey of Business Owners and Self-Employed Persons (2011 or 2015)	228
fem_num_firms	ratio	Women-owned firms	Survey of Business Owners and Self-Employed Persons (2011 or 2015)	49
min_num_firms	ratio	Minority-owned firms	Survey of Business Owners and Self-Employed Persons (2011 or 2015)	564

Appendix 2. Correlation Plots

A2.1. Correlation plot of 2012 dataframe (after log-transformation)



A2.2. Correlation plot of 2016 dataframe (after log-transformation)



Appendix 3. Features selected for from each data set using wrapper.

	LassoCV	RidgeCV	RandomForestRegressor	GradientBoostingRegressor	SVR
2012	population perc_female perc_o18 perc_o65 perc_white_not_hisp median_oo_value monthly_costs_m monthly_costs_nm avg_hh_size perc_1yr_res perc_f_emp_2065 median_income perc_pov num_establishments	population perc_female perc_o18 perc_o65 perc_white_not_hisp median_oo_value monthly_costs_m monthly_costs_nm avg_hh_size perc_1yr_res perc_f_emp_2065 hc_value_1k perc_pov num_establishments	population perc_o18 perc_white perc_black perc_white_not_hisp perc_oo monthly_costs_m monthly_costs_nm num_nonemployers man_value_1k wholesale_value_1k num_establishments annual_payroll min_num_firms	population perc_o18 perc_white perc_black perc_white_not_hisp perc_oo monthly_costs_m monthly_costs_nm num_nonemployers man_value_1k wholesale_value_1k num_establishments annual_payroll min_num_firms	population perc_female perc_o18 perc_o65 perc_hisp perc_white_not_hisp median_oo_value monthly_costs_nm avg_hh_size perc_1yr_res perc_hs perc_f_emp_2065 hc_value_1k perc_pov num_establishments
2016	population perc_o18 perc_o65 perc_white perc_white_not_hisp median_oo_value monthly_costs_nm avg_hh_size perc_hs hc_value_1k median_income perc_pov num_establishments annual_payroll	population perc_o18 perc_o65 perc_white perc_hisp perc_white_not_hisp median_oo_value monthly_costs_nm avg_hh_size perc_hs perc_f_emp_2065 hc_value_1k median_income perc_pov num_establishments annual_payroll	population perc_white perc_black perc_asian perc_white_not_hisp monthly_costs_m monthly_costs_nm perc_bc num_establishments annual_payroll fem_num_firms min_num_firms	perc_white perc_black perc_asian perc_white_not_hisp perc_oo monthly_costs_m monthly_costs_nm perc_bc num_establishments min_num_firms	population perc_female perc_o18 perc_o65 perc_white perc_hisp perc_white_not_hisp median_oo_value monthly_costs_nm avg_hh_size perc_hs perc_f_emp_2065 hc_value_1k median_income perc_pov num_establishments annual_payroll
Combined	population perc_o18 perc_o65 perc_white perc_hisp perc_white_not_hisp median_oo_value monthly_costs_nm avg_hh_size perc_1yr_res perf_f_lang perc_hs perc_f_emp_2065 wholesale_value_1k hc_value_1k median_income perc_pov num_establishments	population perc_female perc_o18 perc_o65 perc_white perc_hisp perc_white_not_hisp median_oo_value monthly_costs_nm avg_hh_size perc_1yr_res perc_f_lang perc_hs perc_f_emp_2065 hc_value_1k median_income perc_pov num_establishments	population perc_white perc_black perc_white_not_hisp median_oo_value monthly_costs_m monthly_costs_nm perc_bc num_establishments annual_payroll min_num_firms	perc_white perc_white_not_hisp perc_oo median_oo_value monthly_costs_m monthly_costs_nm perc_bc wholesale_value_1k num_establishments annual_payroll	population perc_female perc_o18 perc_o65 perc_white perc_hisp perc_white_not_hisp median_oo_value monthly_costs_nm avg_hh_size perc_1yr_res perc_f_lang perc_hs perc_f_emp_2065 hc_value_1k median_income perc_pov num_establishments

Appendix 4. Parameters and most important features for the highest performing model of each algorithm.

	LassoCV	RidgeCV	RandomForestRegressor	GradientBoostingRegressor	SVR
Parameters	'eps': 0.000 'max_iter': 114 'n_alphas': 61	'alphas': array([0.1, 1. , 10.])	'bootstrap': False 'max_depth': 80 'max_features': 'sqrt' 'min_samples_leaf': 2 'min_samples_split': 3 'n_estimators': 1100	learning_rate': 0.03 'max_depth': 10 'max_features': 'sqrt' 'min_samples_leaf': 5 'min_samples_split': 3 'n_estimators': 1000	'C': 10 'gamma': 'scale' 'kernel': 'rbf'
Most important features	Population	One cannot identify the most important coefficients in Ridge Regression	% White, Not Hispanic	% Black	Because the 'rbf' kernel was selected, one cannot identify the most important coefficients in SVR
	% over 65		% White	% White	
	% over 18		Monthly Costs (People w/o mortgage)	Monthly Costs (People w/ mortgage)	
	Median Home Value		% Black	Population	
	Monthly Coss (People w/o mortgage)		Population	% over 18	