131 Final Project S21

Classification of the 2016 general election

David Brackbill, Simranjit Kaur, Joanna Kim, Laila Voss

Introduction

Presidential election predictions are popular practices and in the case of the 2012 election, the predictions proved to be accurate! However, the 2016 election was a huge shock to pollsters expecting a clear Hillary Clinton favorite, although Donald Trump eventually won. What went wrong in 2016 is that the polls had much higher polling error–they may have suffered from selection bias (somewhat fixable with raking) and suffered from a high degree of undecided voters (not fixable) that led poll results to be strongly and misleadingly in favor of Clinton.

Goal

This was a huge lapse in the predictive value of polling, so we wanted to explore further how demographic variables at the county level played into the winner of each county race. We pursued this by conducting analysis on these variables through visualizations, model fitting, and other analysis techniques such as principle component analysis.

Principal component inference

For this project, we used PCA to find which covariates had the greatest influence on determining which candidate would win a certain county. We found PC1, a measure of employment and income, to be the most influential principal component. This suggested that the three largest variables in the first PC were IncomePerCap, ChildPoverty, and Poverty.

Classification approaches

To confirm our findings from the PCA, we constructed both a linear regression model and a quadratic discriminant analysis model. We modeled the probability one candidate wins a county and identified significant associations with state variables using multiple models. The logistic regression model produced superior results, achieving 90.89% predictive accuracy. The model found that White, Citizen, Professional, Service, Production, Drive, Carpool, Employed, PrivateWork, Unemployment were all significantly associated with the county winner.

We also fit a decision tree to the data to predict the winner of each county. We incrementally improved the prediction accuracy of the base model by implementing improvements such as using a threshold tied to the maximum Youden statistic as well as fitting a Random Forest model.

Materials and methods

Datasets

The raw *census* data set was made up of county-level observations in the 2015 American Community Survey. These observations included demographic variables of each county in the United States. The raw *election* data set represented the 2016 US presidential election results on three observational units: county, state and national.

Generally, we pre-processed the data into county-level observations in order to merge the data sets by row on counties. Specifically, the election data was pre-processed by removing rows that did not correspond to any county in the United States. Then, we reformatted the election data to contain only county-level observations. The census data, which was initially on the census tract level, was aggregated to the county level.

This merged data was then used to create state and county maps, along with other useful visuals to capture various relationships between the election candidates and the different covariates. Moreover, we utilized this cleaned data set to build our classification models and explore the different principal components.

Merged data frame used in analysis, first 5 rows and 7 columns

winner	total	Women	White	Citizen	IncomePerCap	Poverty
Donald Trump	24759	51.57	75.79	73.75	24974	12.91
Donald Trump	94261	51.15	83.1	75.69	27317	13.42
Donald Trump	10436	46.17	46.23	76.91	16824	26.51
Donald Trump	8753	46.59	74.5	77.4	18431	16.6
Donald Trump	25442	50.59	87.85	73.38	20532	16.72

Methods

Inference

We utilized ridge charts to visualize the difference between the demographics of the 2 county-level winning candidates, Hillary Clinton and Donald Trump. After carrying out PCA, we examined the proportion of variance explained by each component, as well as the cumulative proportion. We found a sharp drop-off in variance explained after the fourth principal component. In order to select the number of components, we plotted the variances to determine the fewest number of principal components that capture a considerable proportion of the variation and covariation. Ultimately, we decided to stick with the "elbow" value of about four principal components.

GLM

A generalized linear model was used in order to model the probabilities that a major candidate won a county for the 2016 presidential election in order to predict the winning candidate in each county, as well as determine demographic variables that affect the outcome. Discriminant analysis was used to discern if we could do that using linear combinations. The data is split up 80% for training the model and 20% testing the model. The response variable was a factor indicating the winner (Donald Trump 0, Hillary Clinton 1) and the covariates was the census information for the corresponding county.

Using the generalized linear model, the predictors that are statistically significant with the p-value 0.05 will be used for both the generalized linear model and quadratic discriminant analysis model. For both models, the optimal threshold will be calculated using Youden's statistic. Then, the models will be converted to classes using the optimal threshold. These model's classification errors will be compared to find the most accurate model.

Decision tree

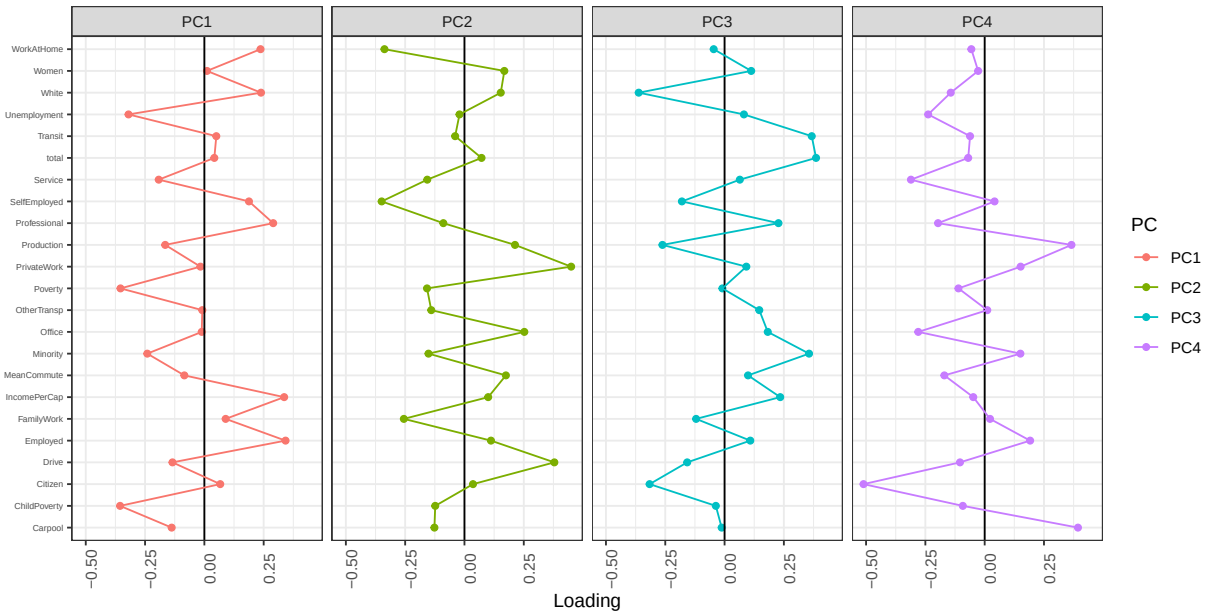
We then implemented a classification decision tree in order to predict the winner in each county using demographic information from the census data, including factors such as income per capita, employment by industry and the prevalence of poverty. To do this, we partitioned the data set, grew and trained the tree and then pruned it using cost-complexity pruning and the deviance loss function. We also ran an iteration in which we used the Gini loss formula. We then compared our predictions to the actual values for each county to examine the classification errors. Finally, we calculated Youden's statistic and implemented a threshold for prediction using this value in an attempt to improve prediction accuracy.

The appeal of a classification decision tree was primarily its intuitive interpretation in this case. Given the level of quantitative detail of the data we were using and the prevalence of studies that indicate the importance of demographic variables such as gender and race on political party preference, it seemed reasonable to assume that using a tree to create decision rules based on these variables could lead to strong predictions.

Finally, as a natural extension of our decision tree analysis, we adopted a random forest method to examine the effect of reducing correlation among the trees and reducing variance. In creating a random forest, we use the ensemble method of bagging on the trees, which aggregates the predictions from models that have been trained on bootstrap samples. The trees are grown through the process of recursive binary splitting on random sunsets of the predictors.

Results

PCA



Our PCA results are detailed below:

PC1 will be **large** when `Unemployment`, `Poverty`, and `ChildPoverty` are **large** and when `Employed`, and `IncomePerCap` are **small**. Given this correlation, we can interpret PC1 as measuring "affluence and employment".

PC2 will be **large** when `White` and `SelfEmployment` are **large** and when `votes` and `Total` are **small**. Given this information, we were unable to find a clear interpretation of PC2.

PC3 will be **large** when `WorkAtHome` and `Minority` are **large** and when `White`, `PrivateWork` and `Drive`are **small**. Given this correlation, we can interpret PC3 as measuring "ethnicity and employment".

PC4 will be **large** when `Production` and `Carpool` are **large** and when `Citizen` is **small**. Given this information, we were unable to find a clear interpretation of PC4.

PC1 was the most influential principal component, capturing almost 25% of the data set's variance. The three largest absolute values of the first principal component were IncomePerCap, ChildPoverty, and Poverty, indicating that these demographic attributes are greatly varied.

GLM and QDA

GLM coefficients chosen by AIC

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-31.03	3.367	-9.216	3.096e-20
White	-0.1352	0.008393	-16.11	2.081e-58
Citizen	0.1575	0.02181	7.221	5.154e-13
Professional	0.2729	0.02742	9.954	2.424e-23
Service	0.382	0.04116	9.281	1.684e-20
Production	0.1312	0.031	4.232	2.314e-05
Drive	-0.1385	0.02183	-6.345	2.223e-10
Carpool	-0.1606	0.04161	-3.859	0.0001138
Employed	0.2175	0.02726	7.98	1.461e-15
PrivateWork	0.1375	0.01633	8.422	3.706e-17
Unemployment	0.2197	0.03582	6.135	8.504e-10

Using the AIC stepwise model selection on the GLM model, the most significant predictors with the p value < 0.05 are White, Citizen, Professional, Service, Production, Drive, Carpool, Employed, PrivateWork, and Unemployment. With Donald Trump encoded as 0 and Hillary Clinton encoded as 1, variable impacts can be seen with the summary of the fitted GLM model. The variable with the largest absolute coefficient is Service.The variables with a larger positive coefficient were Professional, Employed, and Unemployment. Conversely, variables with a negative coefficient were White, Drive, and Carpool.

Errors from GLM and QDA

Table 3: GLM Errors

	Donald Trump	Hillary Clinton
Donald Trump	0.9228	0.07721
Hillary Clinton	0.1786	0.8214

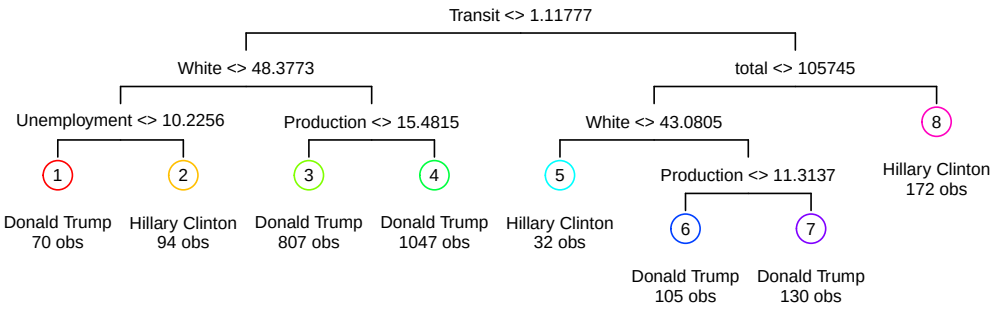
Table 4: QDA Errors

	Donald Trump	Hillary Clinton
Donald Trump	0.8588	0.1412
Hillary Clinton	0.119	0.881

The error rate of the GLM model, with the classes converted using the optimal threshold, was 9.1%. The error rate of the QDA model, with the classes converted using the optimal threshold, was 13.82%. Because the GLM model had lower misclassification rates, it proved to be the more accurate model.

Decision trees

In the tree we trained using the deviance loss function and that we implemented cost-complexity pruning, nine variables were used in construction: the percent of the population that is white; the number of women; the number of citizens; the percent commuting on public transportation; the percent employed in production, transportation and material movement; the number of votes; the percent commuting alone in a car, van or truck; the percent of the population that is a minority; and the percent of the over-age-16 population employed. The total misclassification error rate was 6.3%. We also saw higher accuracy for counties that had voted for Donald Trump, with 93.41% of these counties being correctly classified versus 79.38% for Hillary Clinton.



Then, in order to examine variation in our results based on the randomized seed, we underwent 30 iterations with different seeds. In every iteration, we saw that the classification of counties that voted for Donald Trump had higher accuracy than that of counties that voted for Hillary Clinton. Additionally, three variables were used in every tree: the percent commuting on public transportation, the percent of the population that is white and the number of votes.

Table 5: Average error rate across 30 seeds

Trump	Clinton
0.9708	0.7292

Our next step was to calculate Youden's statistic on the above-mentioned pruned tree in order to find a new threshold to draw predictions along. In doing this, we saw a decrease in the accuracy rate of counties which voted for Donald Trump but an increase in the accuracy rate of counties which voted for Hillary Clinton; the true positive rate for Trump went from 93.41% to 84.50% while the true positive rate for Clinton went from 79.38% to 87.63%.

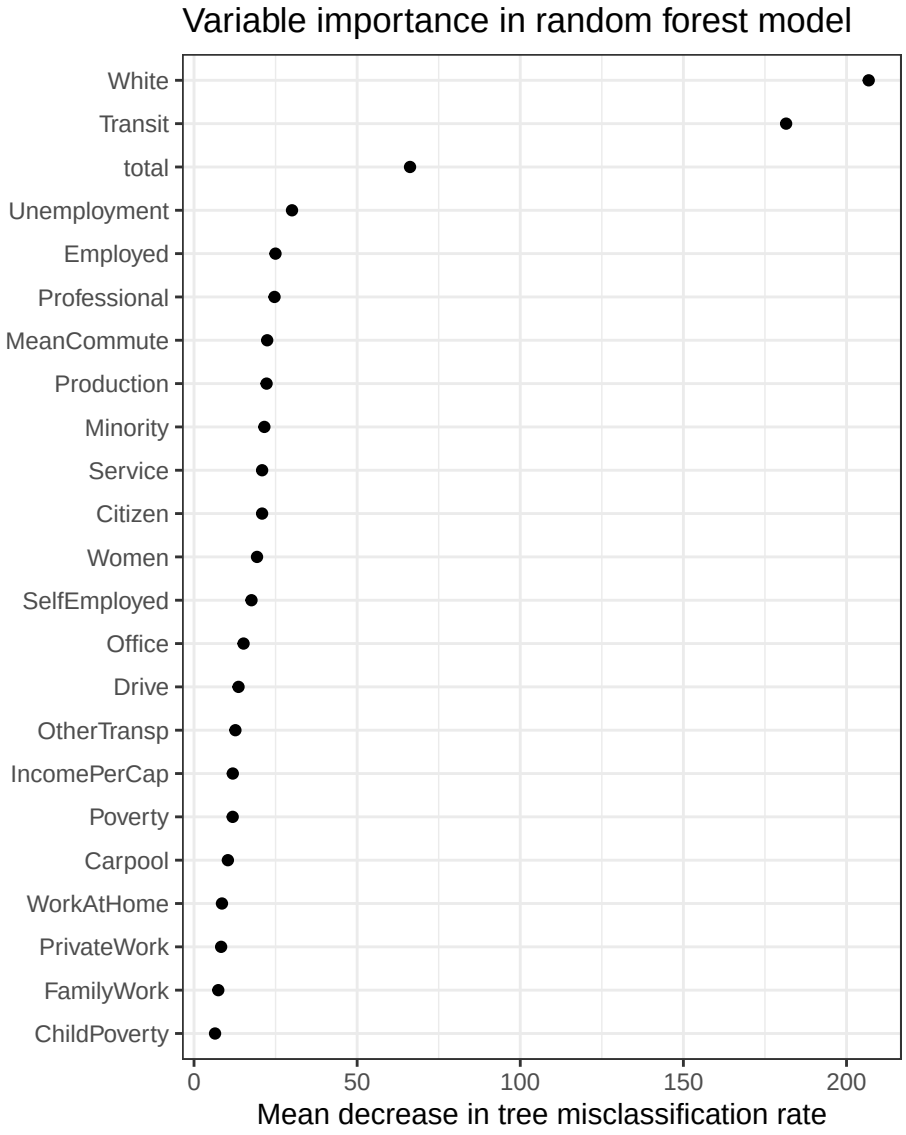
Table 6: Classification success rates, per class, per threshold

	Trump	Clinton
Automatic threshold	0.9341	0.7938
Optimal threshold	0.845	0.8763

In our random forest implementation, we saw a similar trend as in our trees with a lower classification error rate for counties which voted for Donald Trump (2.42%) versus for Hillary Clinton (24.18%). We also examined variable importance in terms of classification accuracy and the Gini index and found the mean decrease in misclassification rate across trees; the most important variables were the same ones that were used in every tree when we ran 30 iterations using different seeding: the percent commuting on public transportation, the percent of the population that is white and the number of votes.

Table 7: Confusion matrix

	Donald Trump	Hillary Clinton	class.error
Donald Trump	2536	70	0.02686
Hillary Clinton	108	356	0.2328



Discussion

Inference through PCA and visualization

After exploring the election data further through visualizations and PCA, we found that race and income played an important role in which candidates won a certain county. Overall we found that income, poverty and race seemed to be valuable factors to take into consideration when examining which covariates have the greatest influence on which candidate would win a certain county.

GLM and QDA takeaways

Out of the 10 predictors used in both the GLM and QDA models, Professional, Service, Employed, and Unemployment had the largest absolute coefficient values, making them the most impactful variables on the county candidate winner. These variables all had positive coefficient values, meaning that counties with higher percentage employed in management, business, science, and arts, percentage employed in service jobs, percentage of employed people older than 16 years, and percentage unemployed, were counties where Hillary Clinton would be more likely to win against Donald Trump.

Conversely, variables with a negative coefficient were White, Drive, and Carpool. This meant that counties with higher percentage of White people, percentage of people driving alone in a car, van, or truck, percentage of people carpooling in a car, van, or truck, were counties where Hillary Clinton would be more likely to not win against Donald Trump.

Some variables in each group of coefficients are contradictory, such as increased employment rate and unemployment rate both making Hillary Clinton's win more likely, while increased independent driving rate and carpool rate both make Hillary Clinton's lost more likely.

Decision tree results and considerations

The variables white, transit, and total were used across all thirty randomized implementations of the decision tree. The significance of these demographic variables on choice of candidate echoes the ex post facto analysis that white people in non-urban areas were significantly more likely to vote for Trump.

The white variable is self-explanatory: the degree of whiteness of a county was a predictor of voting for Trump. Total and transit can be explained as indicators of the population density of a county. Higher numbers of votes and higher numbers of people taking public transit tends to correspond with urban areas, which were more likely to vote for Clinton.

winner	counties
Donald Trump	2606
Hillary Clinton	464

Donald Trump's domination of the rural counties caused some problems for the decision tree. Because there were almost six times as many counties that voted for Trump, Clinton-voting counties were relatively rare, and rare events are by nature hard to classify. As a result, the decision tree with automatic threshold struggled in correctly classifying Clinton counties, which we saw in the average error rates across seeds.

Improvements to the decision tree could be made by implementing a grid search to find hyperparameters that yield the best classification accuracy.

Overall, our strong classification results demonstrate that demographic data is fairly strong at predicting the winner in each county, which is promising for future election cycles.