AngieKay / **Presidential-Election-by-County** Public

☆ **0** stars **0** forks

☆ Star Unwatch ▾

<> **Code** ○ Issues Pull requests ▷ Actions Projects Wiki Security

master had recent pushes about 1 hour ago **Compare & pull request**

master ▾ ...

This branch is 36 commits ahead, 2 commits behind main. Contribute ▾

AngieKay Update README.md … 1 hour ago **36**

View code

Presidential Election Analysis by County

Overview

I am building a model to predict presidential elections at the county level. The results and analysis from this can be used by campaigns for outreach and funding allocation. The model is being trained on the 2016 election results.

Business Understanding

Campaigns can use this more granular analysis to predict how specific counties will vote regardless of how the state as a whole votes.
The main predictors for this model are:

- 2012 election results

- Population of Asian descent
- Population that chose 'other' for race on the census
- Population under 18 and in poverty

After mapping these predictors, it was clear that these counties tend to be clustered in specific areas. Information about these clusters would be useful for campaign outreach.

Data & Methodology

Data was gathered from five different sources.

2016 and 2012 election results: data.world
This dataset includes 2012 and 2016 election results including number of votes for each candidate and total votes from the 2012 election. 2016 election results were used to create the target variable. It also includes density, county type(central/outlying/rural), and land area. This data does not include the state of Alaska.

Alaska 2016 and 2012 election results: thecinyc
This dataset includes the 2012 and 2016 election results for Alaska that were missing from the dataset found at data.world. It does not include the additional information that was found in the dataset above.

Additional Alaska information: wikipedia.com
I pulled missing information about Alaska counties from Wikipedia. This includes density, land area, and county type.

Race: US 2010 Census
I pulled race demographics from the 2010 US census. I kept separate columns indicating people who selected an individual race. The census results get very specific after this and the features were not entirely helpful so there is a summed column of individuals who chose 'two or more races'. Interestingly, race-other was the fourth highest predictor for the final model. In the future I would like to do a deeper analysis on that feature.

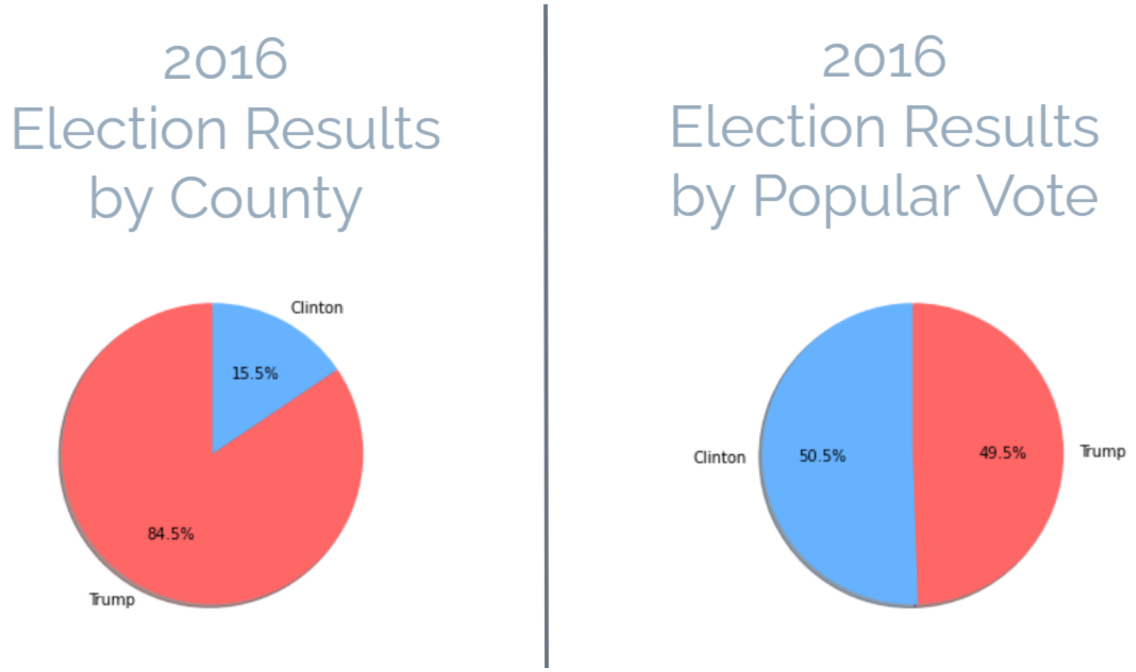
Income: US 2010 Census
I pulled income information from the 2010 US census. This includes total poverty, poverty under the age of 18, and median household income.

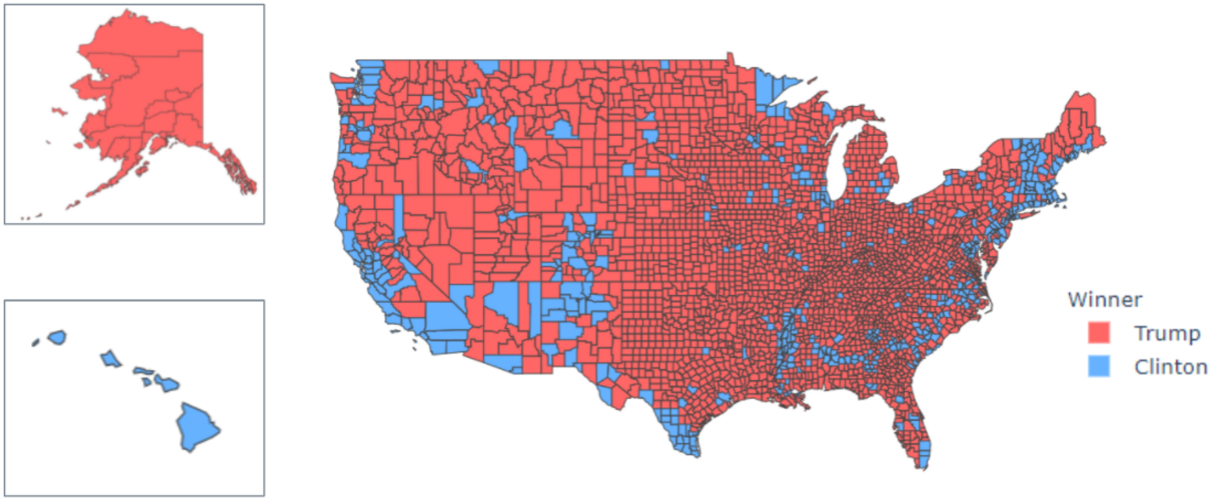
After preprocessing, joining, and one hot encoding two categorical columns, the final dataset has 71 features.

EDA

Target

Donald Trump won 84.5% of counties in 2016 and Hillary Clinton won 15.5%. I've included a pie chart of the popular vote for reference. While Donald Trump won the vast majority of counties, he did not win the popular vote.

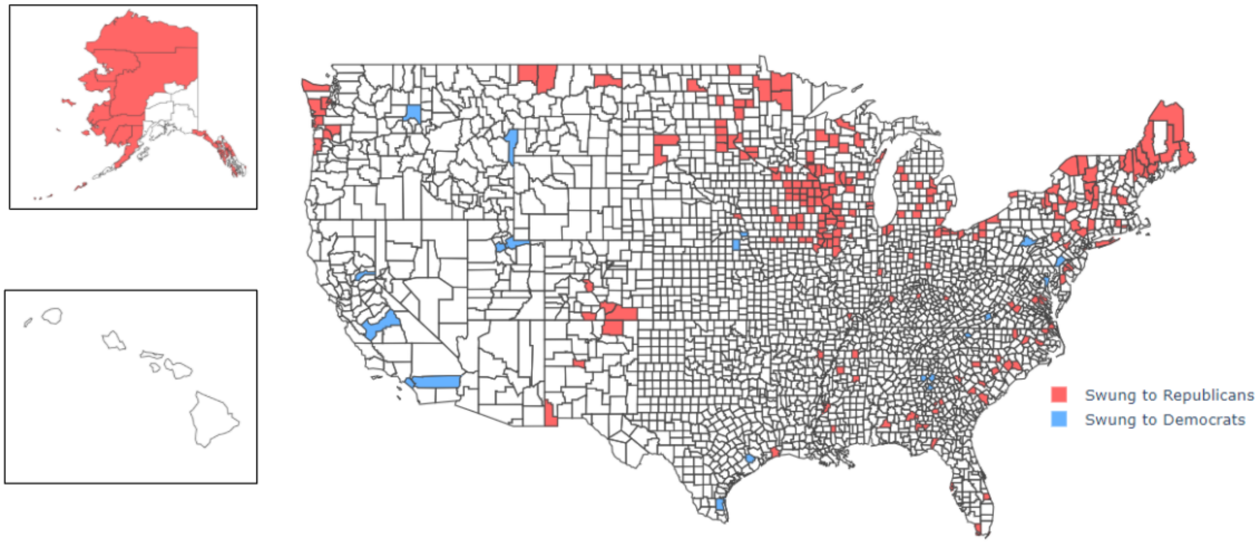




Previous Election

Counties that swung between 2012 and 2016

Of the 3141 counties in the US, 253 of them swung in 2016. 92% of those 253 swung Republican.



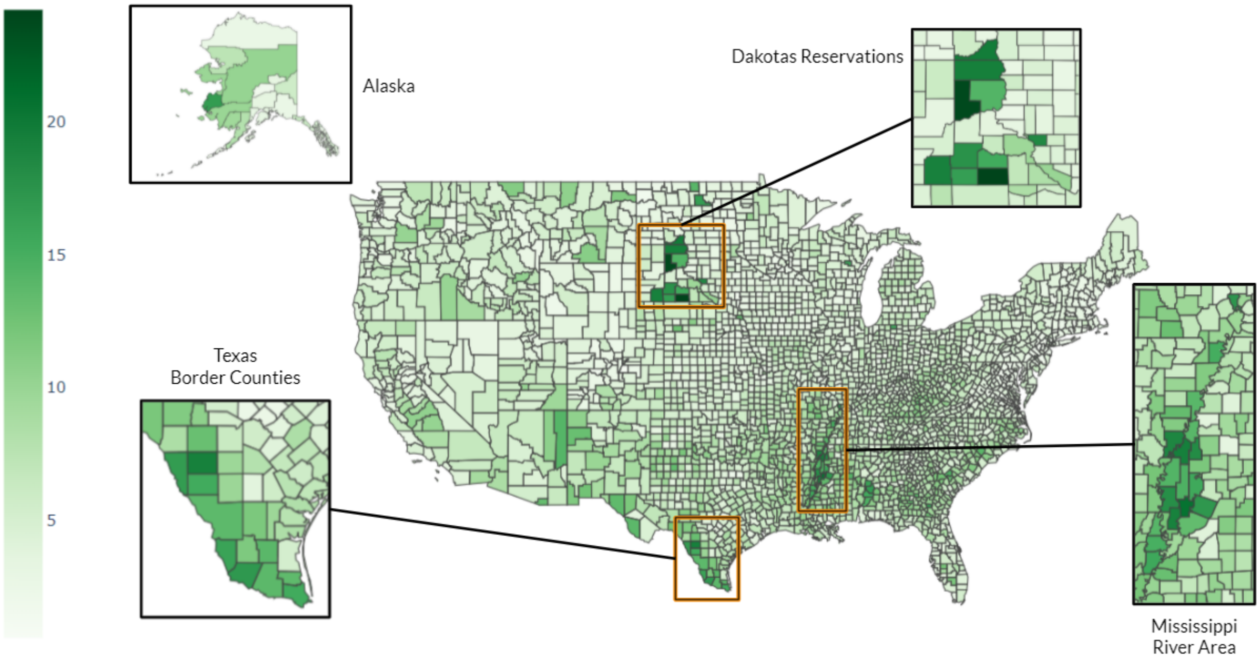
Areas that would be interesting to investigate:

- Maine
- Alaska
- Iowa, Wisconsin, Illinois cluster
- Southern Colorado and northern New Mexico

Childhood Poverty

Childhood poverty rate

This model includes childhood poverty as a raw number, not as a rate. This map explores the rate, not the raw number.



The three main clusters featured in this map all have a very interesting history.

Texas/Mexico Border

This region was heavily involved in the Texas Revolution and most of these counties are named after men involved in that revolution. These counties are over 80% Hispanic.

Dakotas Reservations

This cluster encompasses the Rosebud Indian Reservation, the Standing Rock Indian Reservation, the Pine Ridge Reservation, the Spirit Lake Indian Reservation, the Cheyenne River Indian Reservation, and the Crow Creek Indian Reservation. The Wounded Knee Occupation by the American Indian Movement happened on the Pine Ridge Reservation. The Standing Rock pipeline protests took place on the Standing Rock Reservation. The population of these counties are all over 50% Native American and the majority of them are over 70% Native American.

Mississippi Delta and Surrounding Counties

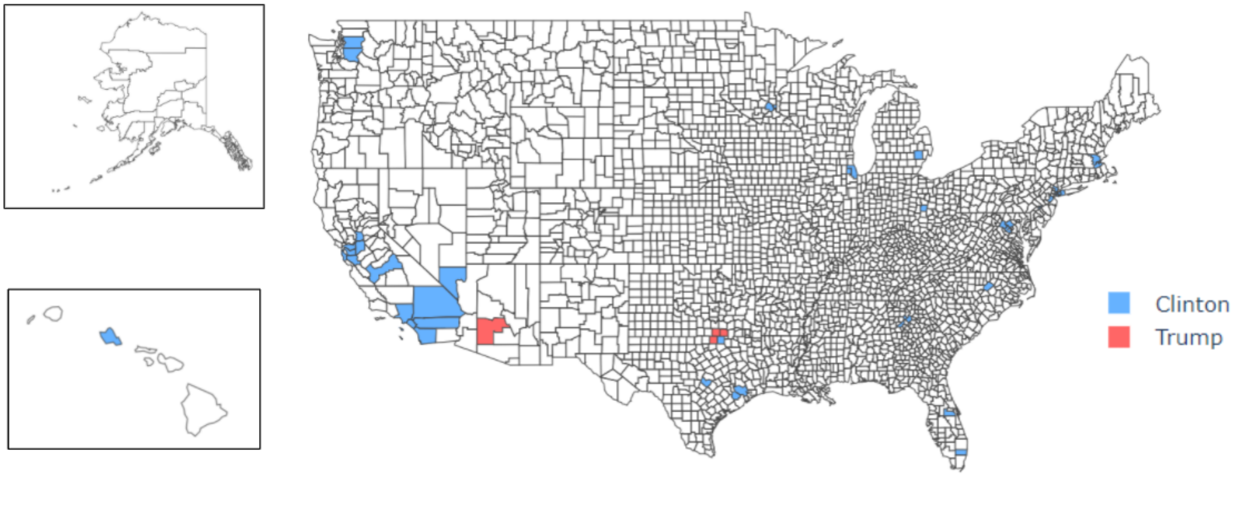
Martin Luther King Jr. originally wanted the Poor People's Campaign to start in Quitman County because of the intense and visible economic disparity there. From 1877 to 1950, there were 48 documented lynchings of African Americans in Leflore County. Phillips County has the highest number of lynchings in US history. The population of these counties are predominantly African American.

The information above was all gathered from Wikipedia.com.

People of Asian Descent

50 counties with highest number of people of Asian descent

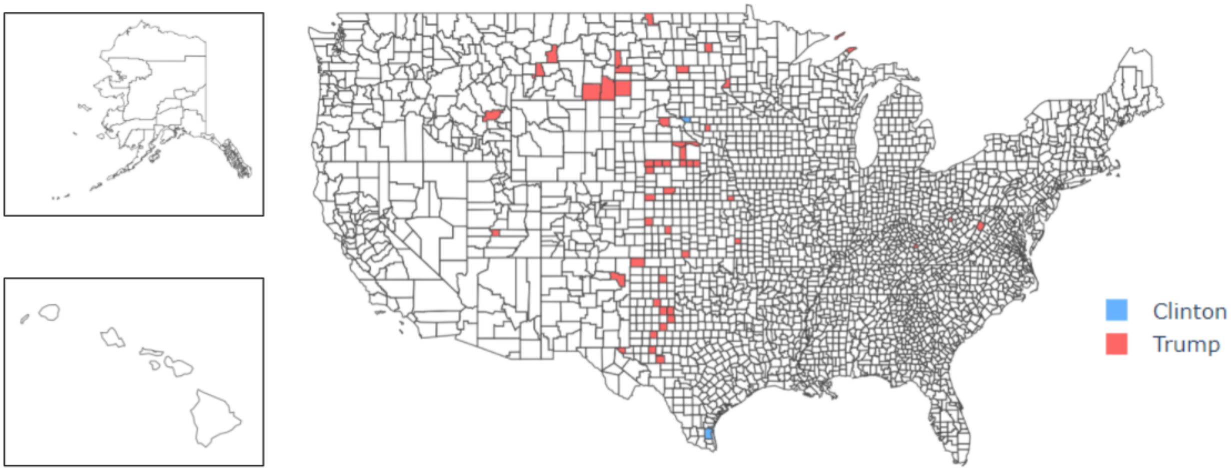
Of these 50 counties, 46 voted for Hillary Clinton.



50 counties with lowest number of people of Asian descent

Of these 50 counties, 48 voted for Donald Trump.

37 of these counties have 0 people of Asian descent and the remaining 13 have 1 person of Asian descent.

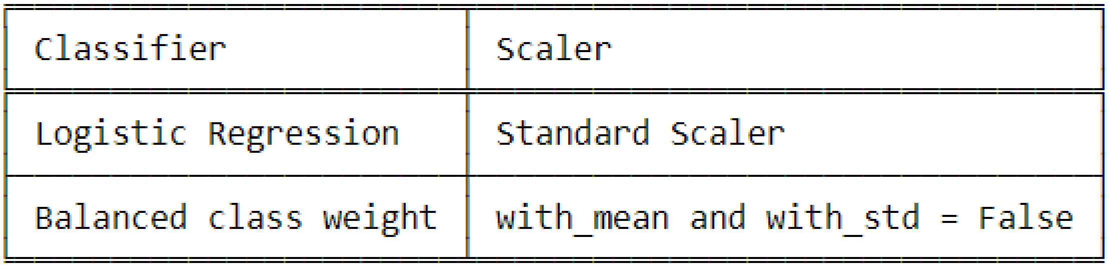


Models

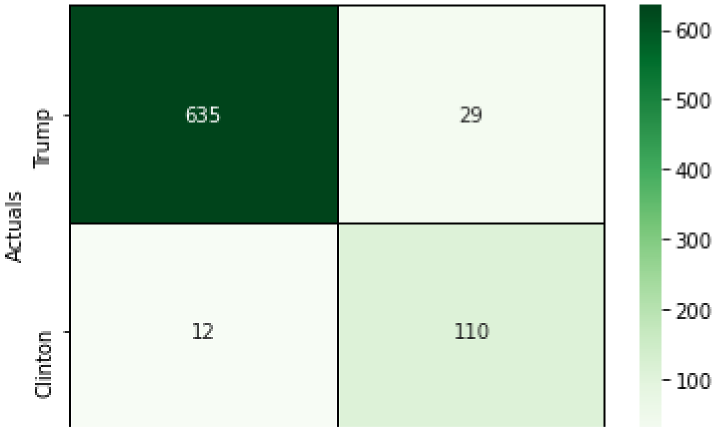
My first simple model was a c-support vector classification with a standard scaler. It had a balanced accuracy score of 76.8%. I spent some time tuning a linearSVC model before switching to logistic regression. I used different scalers, no scaler, SMOTE, and tried different hyperparameters for the classifier and scaler. I one hot encoded the categorical features and used grid search to try different classifiers, scalers, and parameters. Cross validation was used and a test set was held until the end.

Final Model

The final model uses logistic regression with default parameters other than class weight which was changed to 'balanced'. It uses a standard scaler with 'with_mean' and 'with_std' set to False and has a 92.9% balanced accuracy score.



Classifier	Scaler
Logistic Regression	Standard Scaler
Balanced class weight	with_mean and with_std = False



README.md

Conclusions

The largest predictors for predicting presidential elections at the county level are:

- Previous election results
- Population of Asian descent
- Population that chose 'other' on the census
- Childhood poverty

After mapping it was obvious that these things come in clusters or pockets throughout the country. The 2016 election saw many counties swing Republican and very few swing Democrat. The extremes of population of Asian descent are very interesting and vote very predictably. Regions with high childhood poverty rates tend to be clustered in areas with a history of protests, severe poverty, and war.

Future Work

- Include poverty rates rather than just raw numbers.

- Exclude some features and rerun the model to see how it performs. I'm thinking there are some unnecessary features included at this point.
- Further analysis of other key predictors, especially 'race-other' and density.
- Explore and map the counties that the model classified incorrectly. Are they the same counties that swung from 2012 to 2016?
- Do additional model iterations.

For More Information

See the full analysis in these jupyter notebooks: Preprocessing, EDA, Modeling

Angie Rincon
Data Scientist
Github
LinkedIn
angiekay.rincon@gmail.com

Releases

No releases published
Create a new release

Packages

No packages published
Publish your first package