The Divided States of America

Dmitry Mikhaylov and Keita Miyaki

https://github.com/keita-dc/FlatironProject5

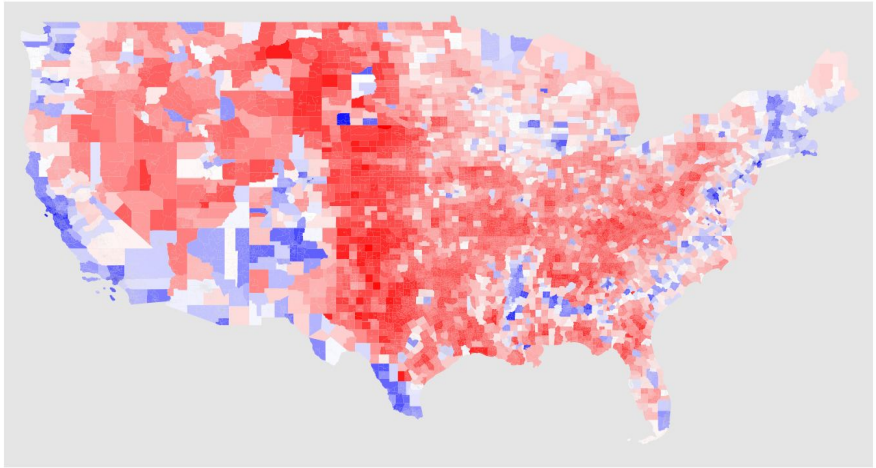
Hmm...

Our questions

What did we ask?

1. Can census data tell the winner of 2016 presidential election at a county level?

2. What features matter most?



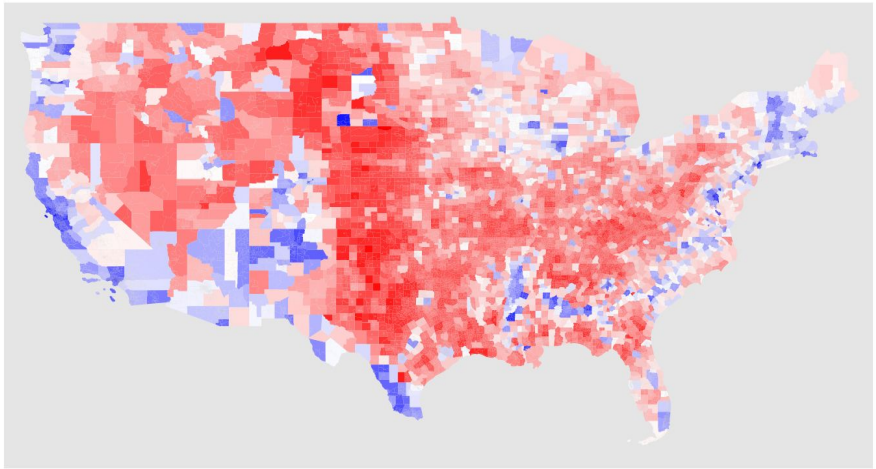
Aha!

Our discoveries

What did we learn?

1. The nation is so divided that most models can predict the county-level results well

2. Race and Urbanization (housing, transportation, etc) do matter most



Our Dataset

US Census

American Community Survey 5-Year Data (2017)

- 260 features after cleaning
- 3,112 countries (excluding Puerto Rico and Alaska)

Our Features

ACS “Data Profiles”

Data Profiles contain broad social, economic, housing, and demographic information. The data are presented as population counts and percentages. There are over 1,000 variables in this dataset.

https://www.census.gov/data/developers/data-sets/acs-5year.html

Our Features Data Cleaning

ACS “Data Profiles”

- Only percentage point estimates
- Only features provide county-level data
- Only features with range between 0 and 100

Our Feature Examples

Demography

- Race
- Age
- Sex

Urbanization

- Apartments/house
- Means of transportation

Income

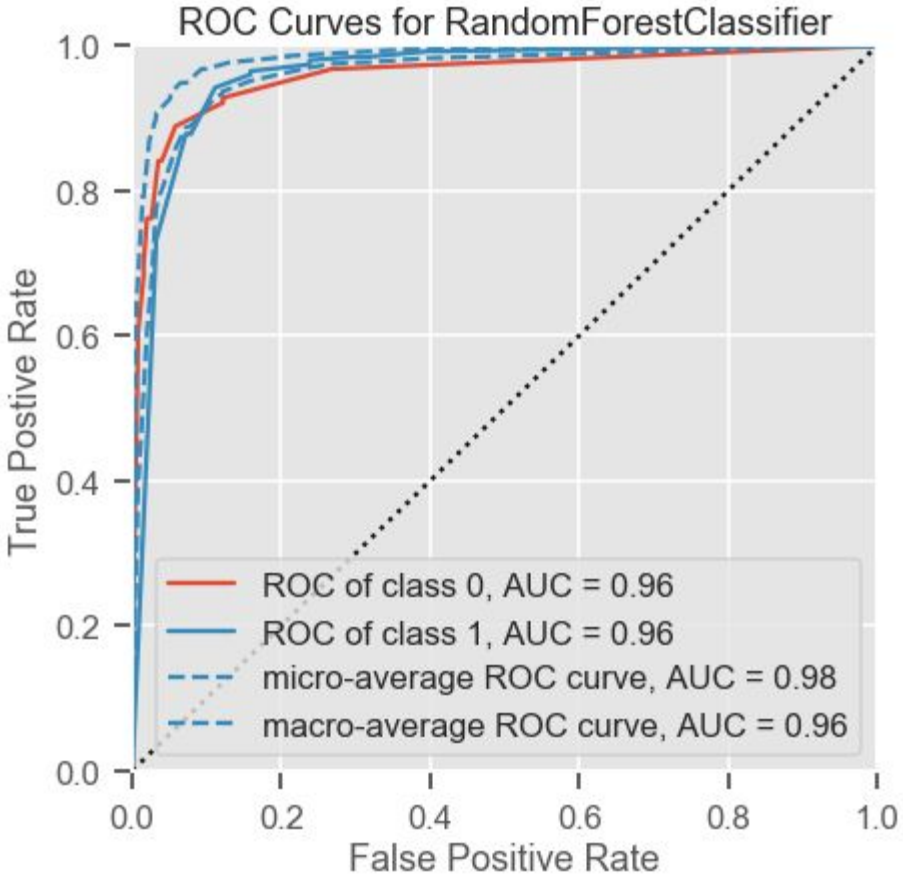
- Income range

Our Analysis

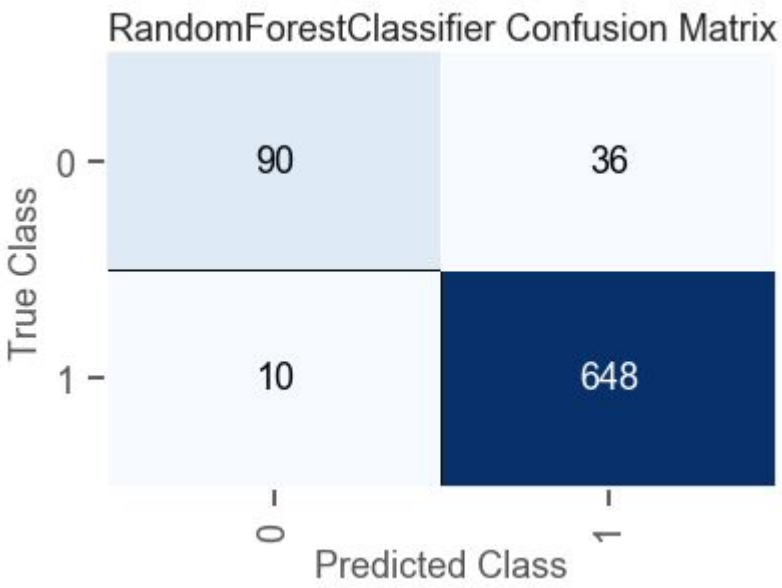
Census ACS Data
- 260 Features
- 3,112 Counties

Models
- Random Forest
- SVM
- XGBoost

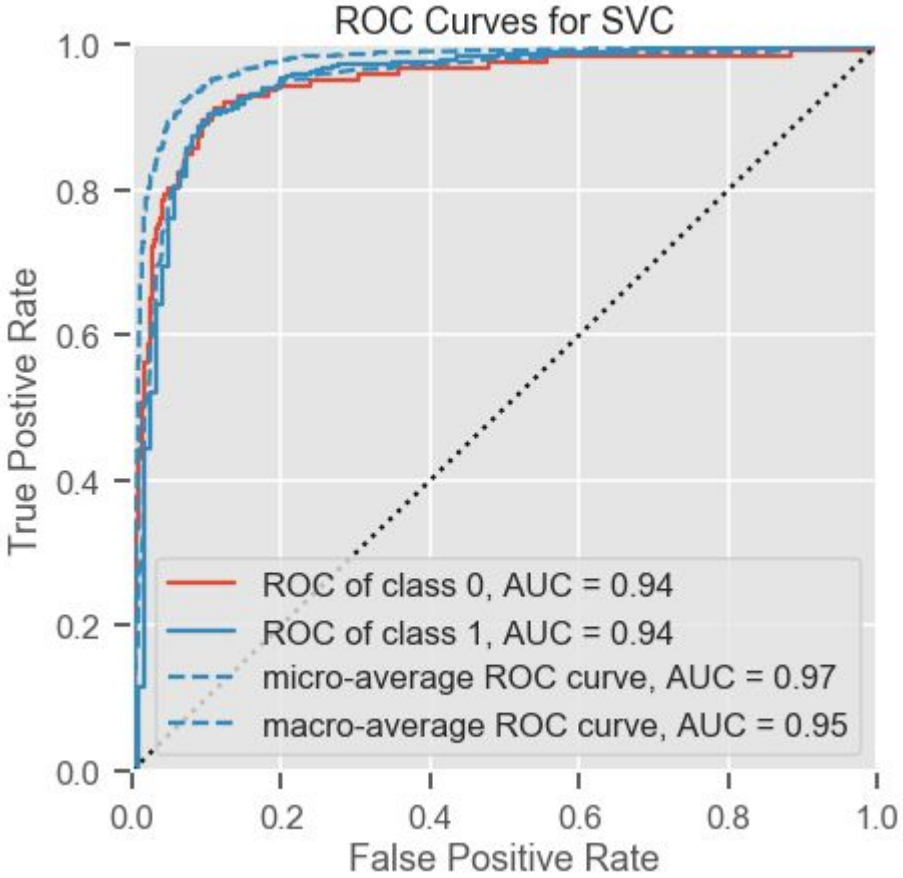
Candidates
- Hillary (0)
- Trump (1)



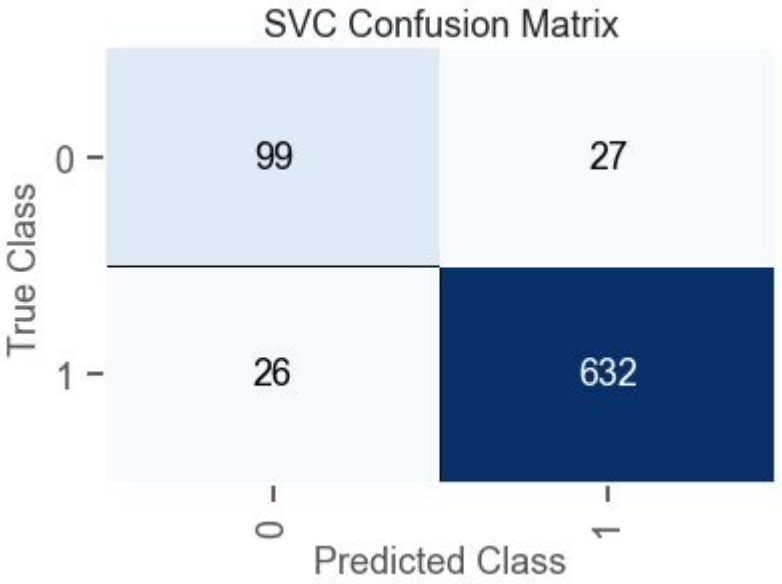
Random Forest



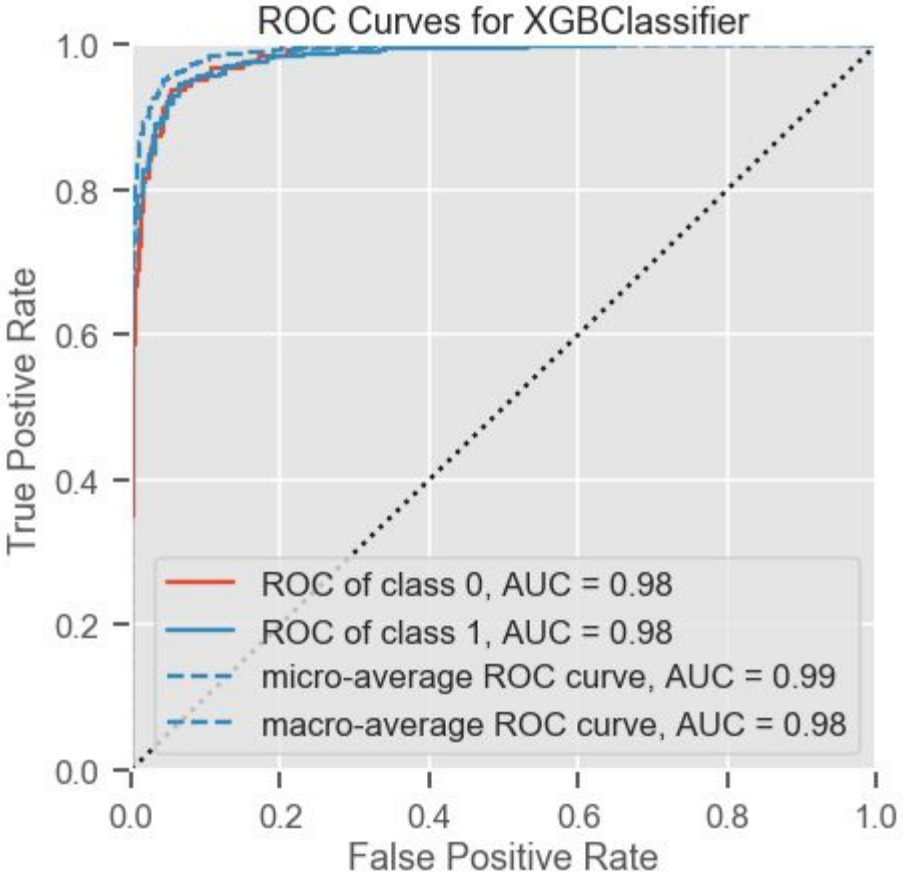
Random Forest



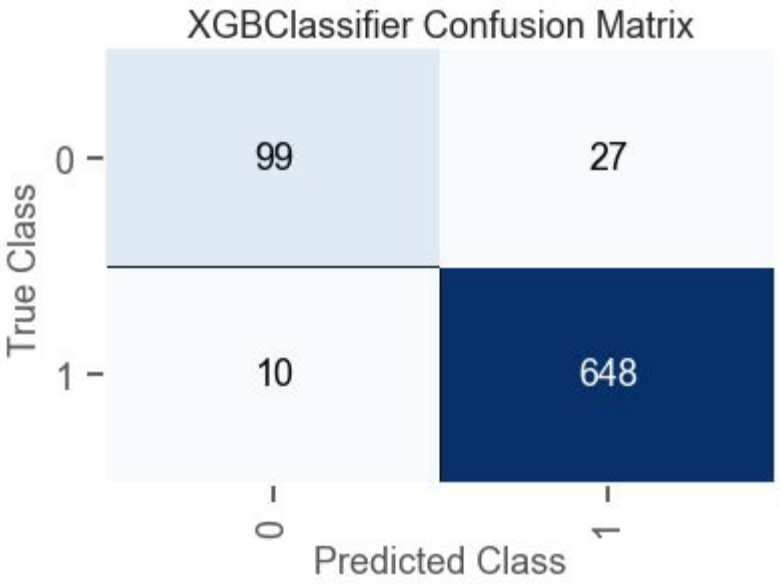
Support Vector Machine



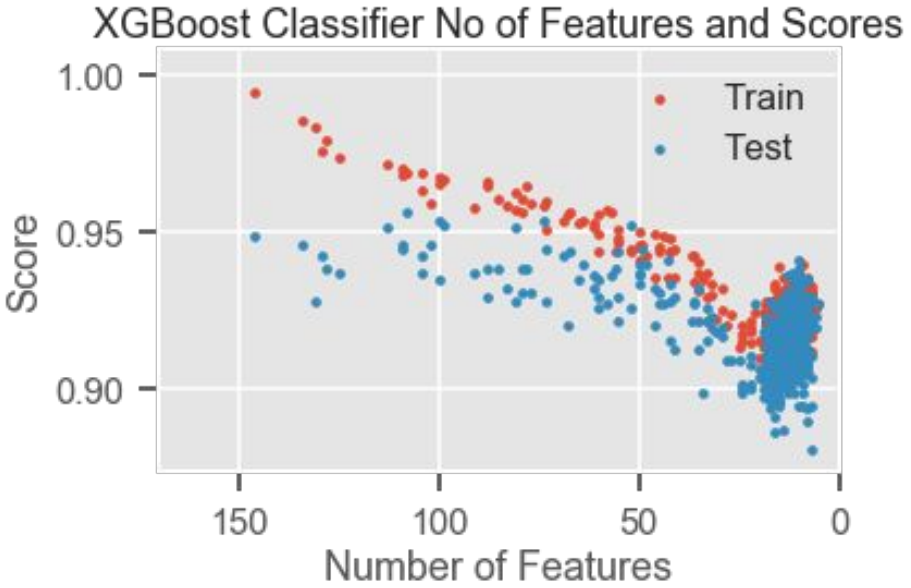
Support Vector Machine


ROC Curves for XGBClassifier
True Postive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC of class 0, AUC = 0.98
ROC of class 1, AUC = 0.98
micro-average ROC curve, AUC = 0.99
macro-average ROC curve, AUC = 0.98

XGBoost

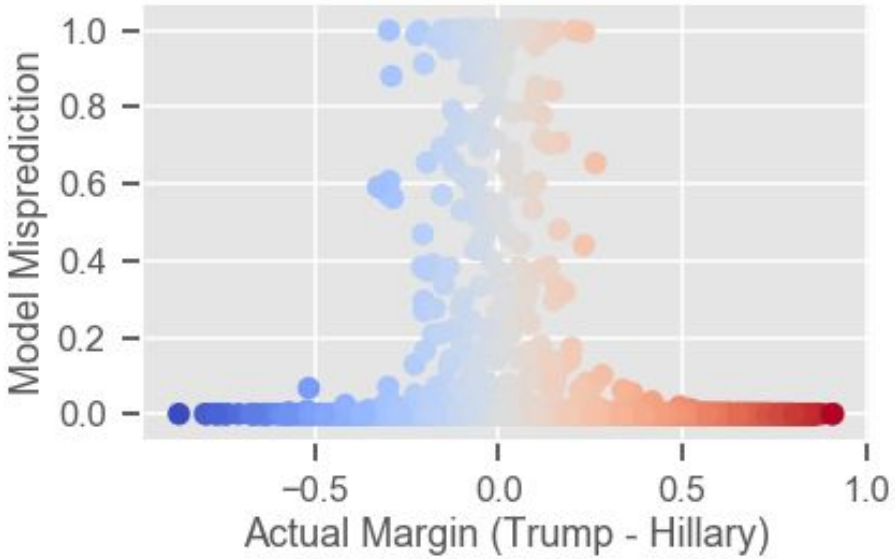

XGBClassifier Confusion Matrix
True Class
0
1
99
27
10
648
0
1
Predicted Class

XGBoost



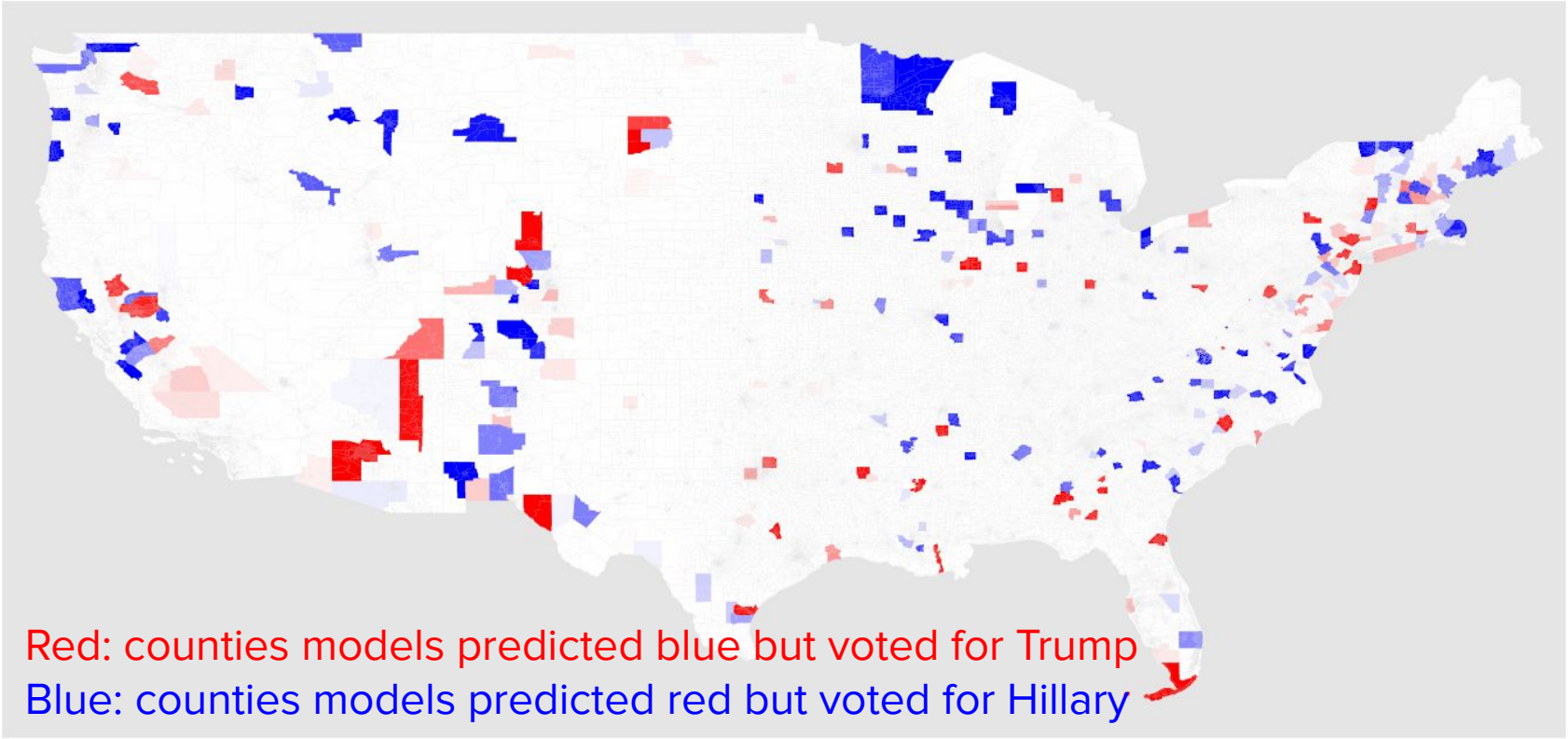
Feature Selection in XGBoost by Regularization

Important Features



Margin vs Misprediction

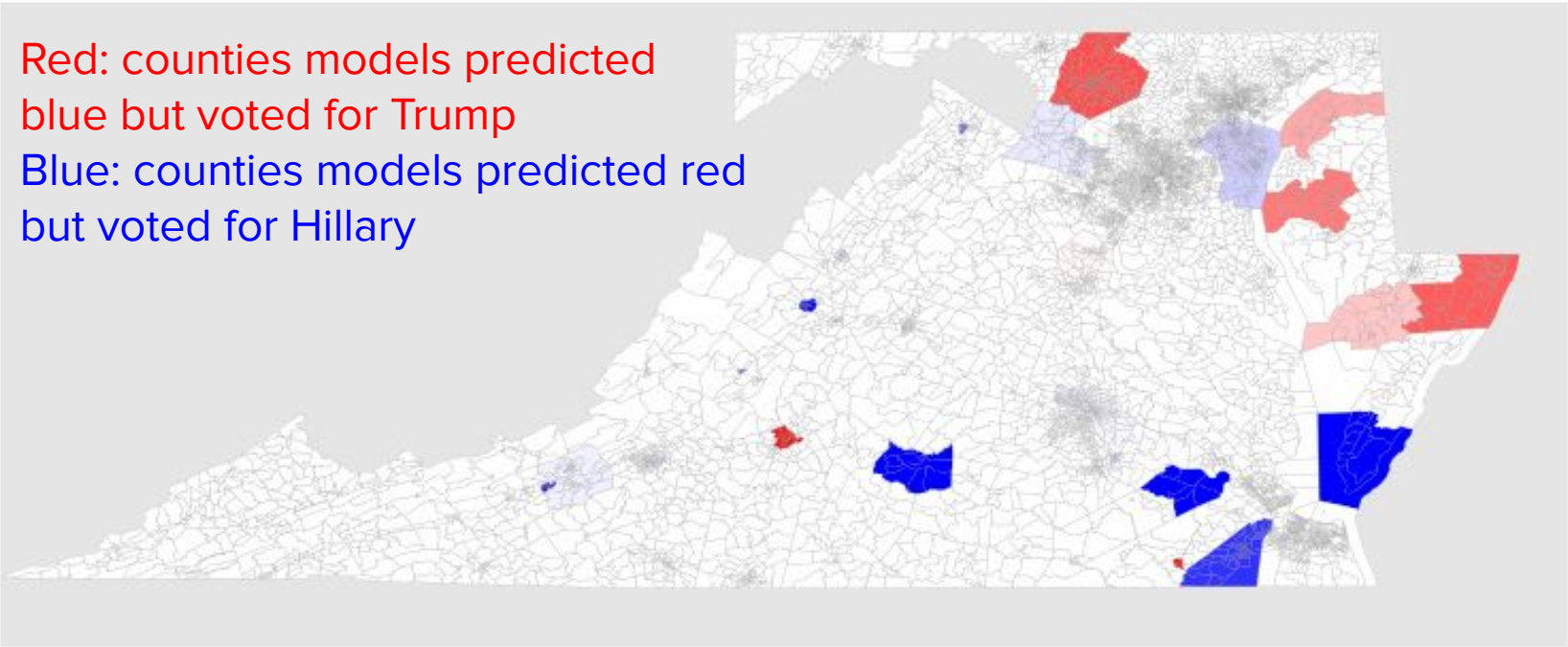
Confusion Map - Where XGB Mispredicted



Confusion Map - Where XGB Mispredicted

Red: counties models predicted blue but voted for Trump

Blue: counties models predicted red but voted for Hillary



Thank You!

&

Go Vote!