DATA SCIENCE FOR U.S. GENERAL ELECTION OUTCOMES

FRANK SCHIEBER

DECEMBER 20, 2017

Project Thesis

Can U.S. Census Bureau features and data be used to predict U.S. General Election votes?

Data Acquisition

"The leading source of quality data about the nation's people and economy"

https://www.census.gov/

Data Acquisition

AMERICAN COMMUNITY SURVEY

U.S. CENSUS BUREAU

"Ongoing survey that provides vital information on a yearly basis about our nation and its people."

https://www.census.gov/programs-surveys/acs.html

Data Acquisition

AMERICAN COMMUNITY SURVEY

U.S. CENSUS BUREAU

"Information from the survey generates data that help determine how more than **$675 billion in federal and state funds** are distributed each year."

https://www.census.gov/programs-surveys/acs.html

Data Acquisition

The American Presidency Project™

"Non-profit and non-partisan, the leading source of presidential documents on the internet"

http://www.presidency.ucsb.edu/elections.php

Data Acquisition

- 50 U.S. states and District of Columbia only
  - Puerto Rico U.S. territory and prohibited from voting
- U.S. Census data from individual general election years 2008, 2012 and 2016
  - Readily available by year from 2007 to present
- Toughest challenge: data munging!

Data Acquisition

- U.S. Census subjects
  - Social
  - Economic
  - Housing
  - Demographic
- U.S. Census subject elements
  - Estimate (520 features in 2016)
  - ~~Margin of error~~
  - ~~Percent~~
  - ~~Percent margin of error~~

= ‘Democrat Votes’

= ‘Republican Votes’

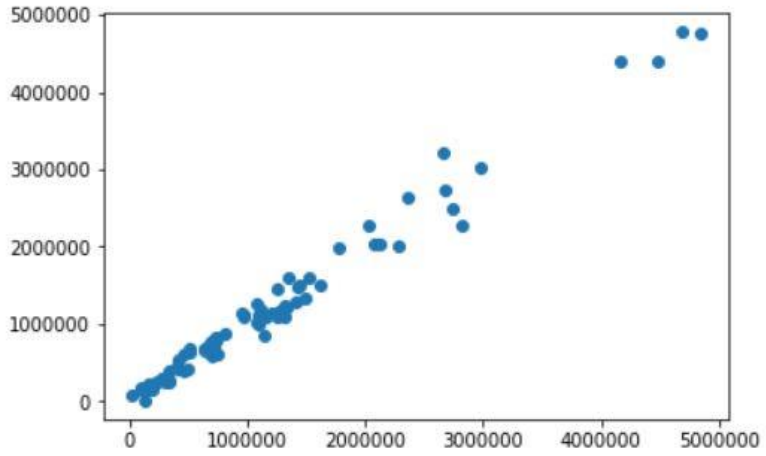
Model Selection

- Linear regression model
  - Train-test-split (TTS)
  - Cross-validation
  - Scatterplots

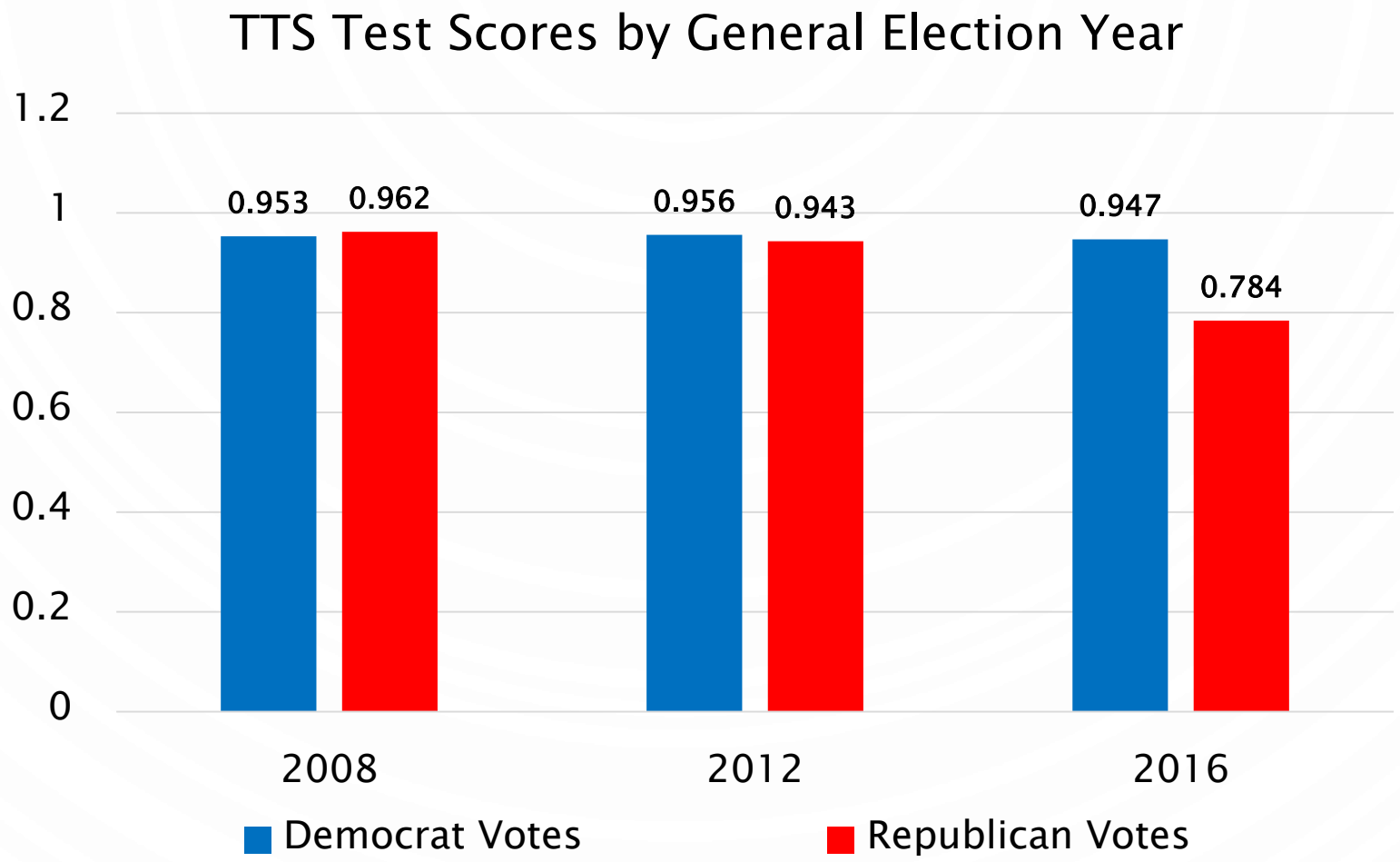
```
In [28]: 1 slrRep.score(XRep_test, yRep_test)
Out[28]: 0.97982268628086067

In [29]: 1 predsRep = slrRep.predict(XRep_test)

In [30]: 1 plt.scatter(yRep_test, predsRep)
Out[30]: <matplotlib.collections.PathCollection at 0x26cd258d8d0>
```



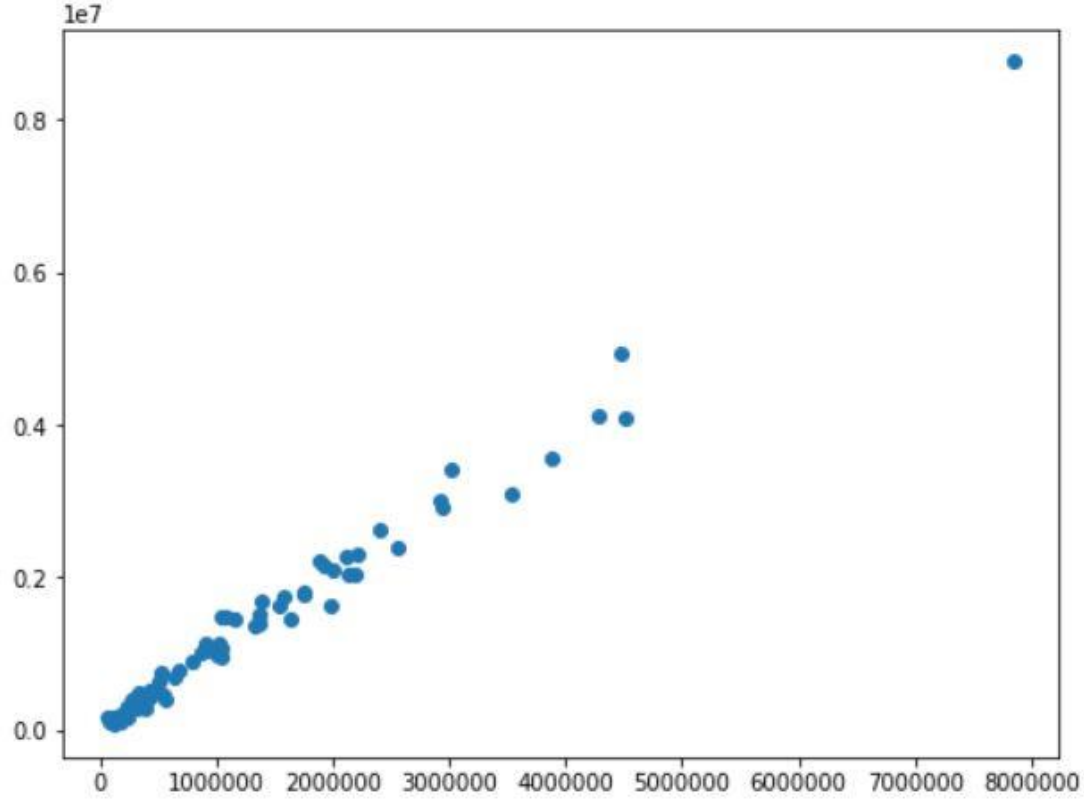
Chronological Approach



Chronological Approach

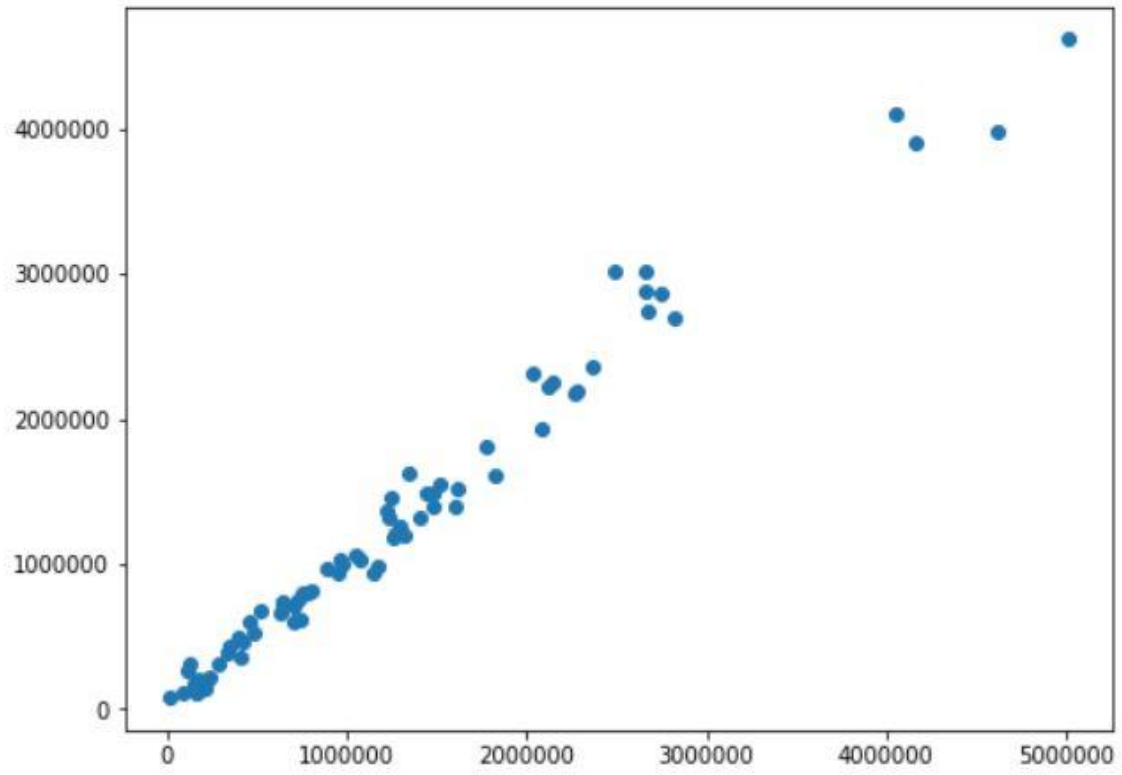
- Full estimated feature set
- Estimated features lower predictive value for Republican votes than Democrat votes, especially in 2016

Compiled Approach: Democrat



Test score = 0.977

Compiled Approach: Republican

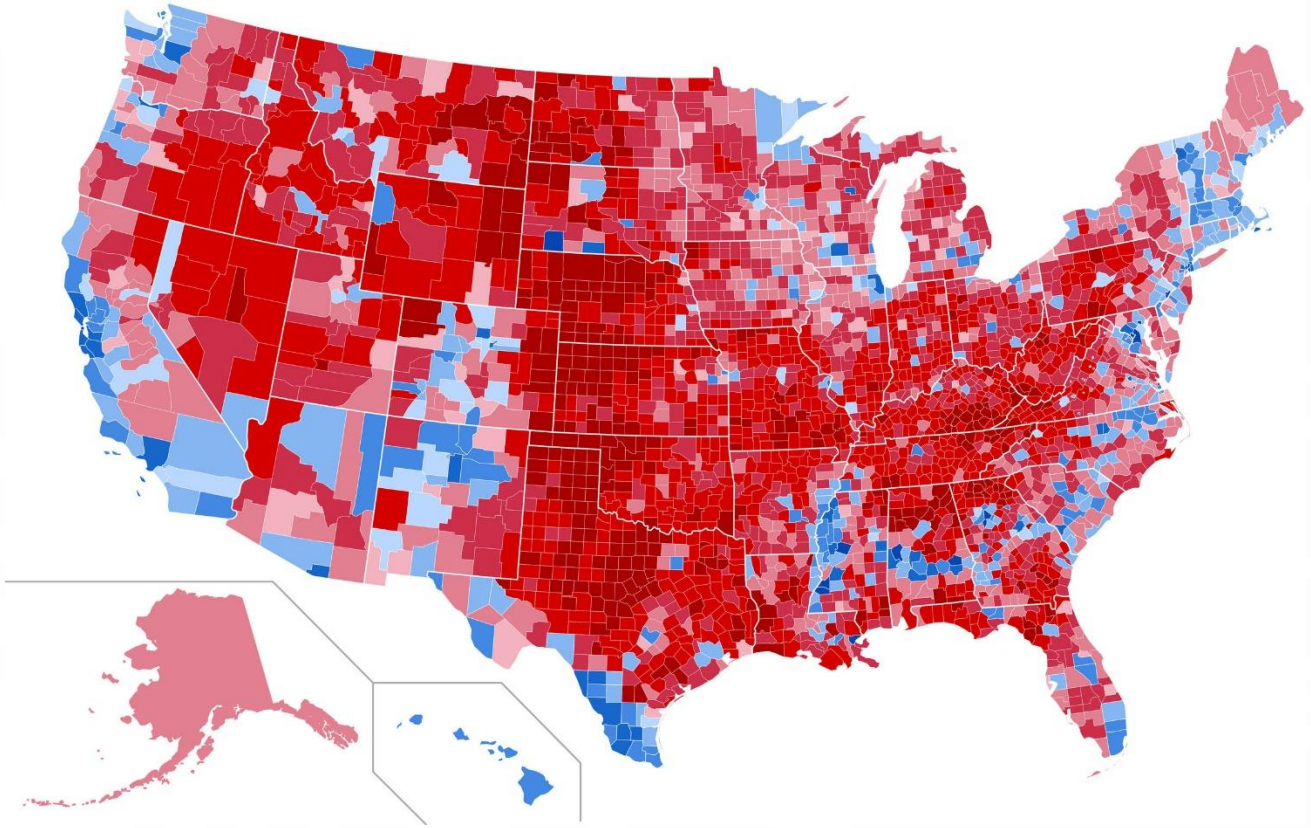


Test score = 0.979

Compiled Approach

- Estimated Social subject feature subset
  - Households by type
  - Relationship
  - Marital status
- Estimated Social features *slightly* higher predictive value for Republican votes than Democrat votes
- Scatterplots highly linear and positive

Conclusion



- U.S. Census data, specifically from the American Community Survey, is **highly effective** in predicting state voting values


made with
Biteable.com
Further Exploration
made with
Biteable.com

FRANK SCHIEBER

678.773.4069

frankschieber@att.net