Classification and interpretation

2016 presidential elections

Data sources

2016 presidential election results, by county

US Census data, 2008 - 2014, by county:

- 51 numerical quantities, e.g. 'PST045214', 'Population, 2014 estimate', 'INC110213', 'Median household income, 2009-2013'

County level data is not representative ...

- Los Angeles County, (pop. 10 million) =? Loving County, TX, (pop. 86)
- In the 2016 election, Republicans won 84 % of the counties while losing the popular vote.

… or relevant to the outcome ...

- Most states' electoral votes are winner-take-all.

… but could be useful for campaigns

Tools and methods

Logistic regression, predicting winner of each county with census attributes as predictors

AUC is the primary metric, evaluated on the hold-out set

Training uses oversampling of the minority class (Democrats win) while the hold-out set is not oversampled

scikit learn

SEA BORN

Bokeh

Single features can score fairly well

Logistic Regression on the best feature, ‘RHI125214’, gives AUC of 0.80

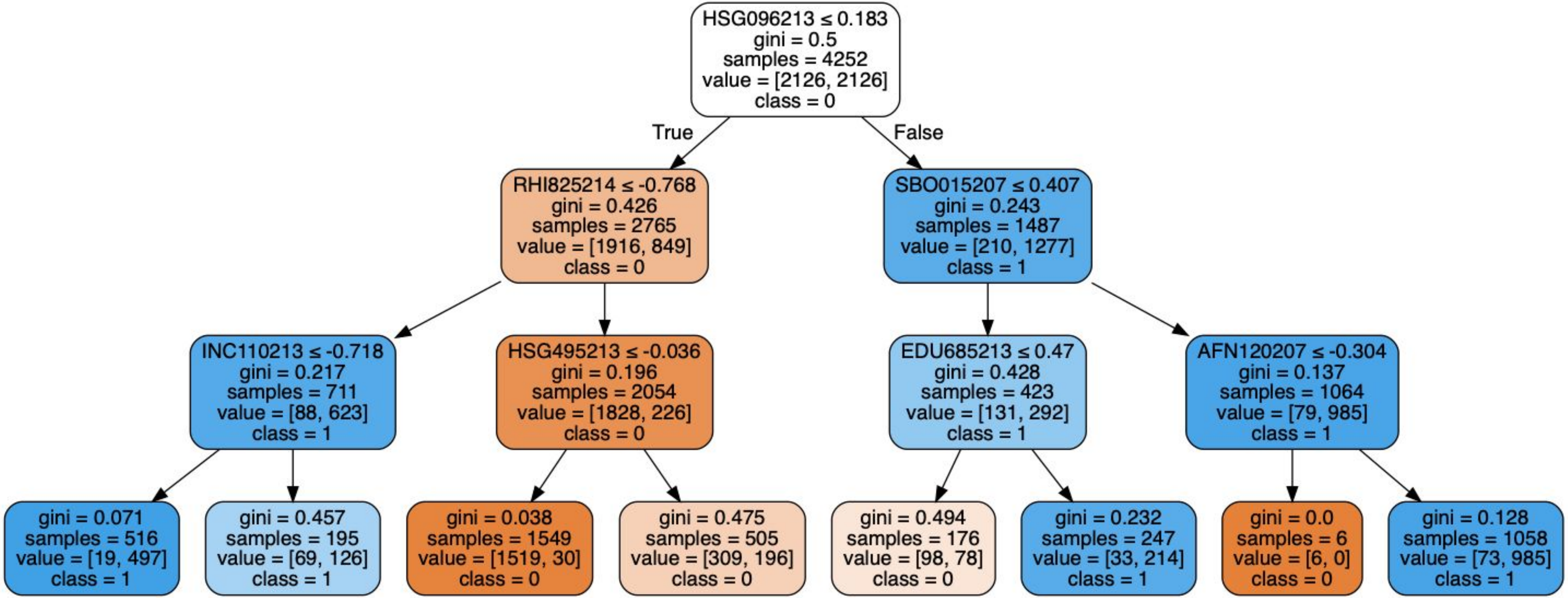
Complex models score well

RFECV (Recursive Feature Elimination - CV) with Logistic Regression includes 46 of the 51 census features, AUC score of 0.97 on the hold-out set

A RandomForestClassifier is 93.3 percent accurate on the holdout set.

Can we come up with something simpler that still scores well?

What is “interpretable”? Maybe a small tree



87 percent accuracy on hold-out set

Choosing the top 5 features with RFE …

AUC score of 0.95 in holdout set with Logistic Regression on the top 5 features

But this score is pretty close to the complex model AUC of 0.97.

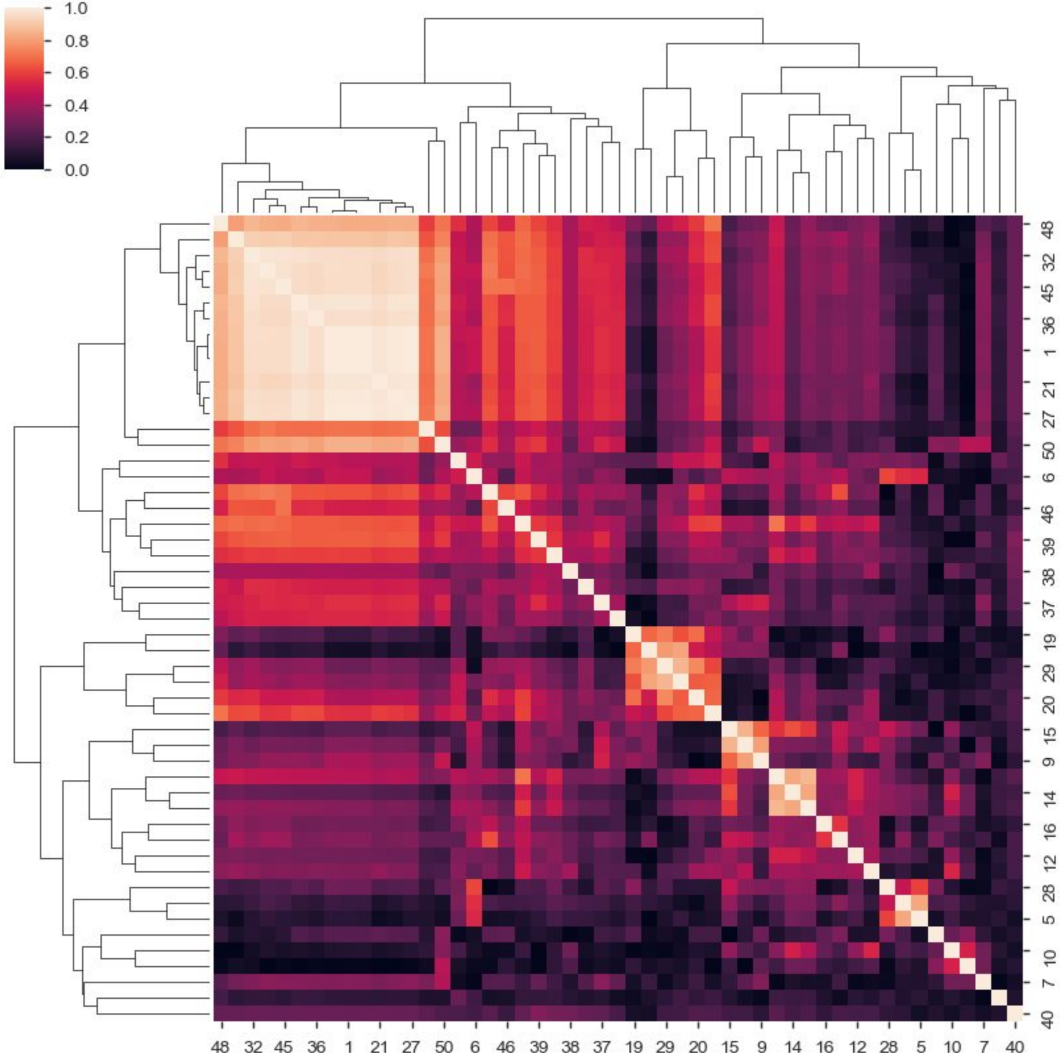
Choosing an additional 3 interaction features gives an AUC of 0.96, closer still.

Feature correlations and clustering

Cluster absolute values of correlation coefficients, see how regressors form groups

This uses the regressors only, not the target

Uses scipy.cluster.hierarchy and sns.clustermap



Extract clusters from hierarchy

Regressors grouped into six clusters.

The one with the strongest correlations consists of features that scale with population

```
('PST045214', 'Population, 2014 estimate')
('PST040210', 'Population, 2010 (April 1) estimates base')
('POP010210', 'Population, 2010')
('VET605213', 'Veterans, 2009-2013')
('HSG010214', 'Housing units, 2014')
('HSD410213', 'Households, 2009-2013')
('BZA010213', 'Private nonfarm establishments, 2013')
('BZA110213', 'Private nonfarm employment,  2013')
('NES010213', 'Nonemployer establishments, 2013')
('SBO001207', 'Total number of firms, 2007')
('MAN450207', 'Manufacturers shipments, 2007 ($1,000)')
('RTN130207', 'Retail sales, 2007 ($1,000)')
('AFN120207', 'Accommodation and food services sales, 2007 ($1,000)')
('BPS030214', 'Building permits, 2014')
('POP060210', 'Population per square mile, 2010')
```

Other clusters consist roughly of ethnic, economic, or educational features

('EDU635213', 'High school graduate or higher, percent of persons age 25+, 2009-2013')
('EDU685213', "Bachelor's degree or higher, percent of persons age 25+, 2009-2013")
('HSG495213', 'Median value of owner-occupied housing units, 2009-2013')
('INC910213', 'Per capita money income in past 12 months (2013 dollars), 2009-2013')
('INC110213', 'Median household income, 2009-2013')
('PVY020213', 'Persons below poverty level, percent, 2009-2013')

In this case, using clusters to directly assist in selecting features was not helpful

For example, selecting the best feature from each cluster did not give better results than just selecting the top features with RFE.

But Clusters may still offer some insight

We can anticipate or understand the groups of regressors picked out by the clustering. This may be helpful for interpretation.

Thank you

Appendix

Further work

Compare classification results with regression on percent dem vs. rep

Explore Box Cox or power transformations of regressors

Weight counties by population to explore population trends

Group counties by state for relevance to electoral vote allocation

AUC - is this a satisfactory metric? Look into automatic selection of a decision threshold

The top 5 features for predicting county outcome:

```
('RHI825214', 'White alone, not Hispanic or Latino, percent, 2014')
('EDU685213', "Bachelor's degree or higher, percent of persons age 25+, 2009-2013")
('HSG495213', 'Median value of owner-occupied housing units, 2009-2013')
('HSD410213', 'Households, 2009-2013')
('INC110213', 'Median household income, 2009-2013')
```

Top five features plus interactions

```
Num Features: 8
1 :  [1 0 0 0 0 0 0 0 0 0] : AFN120207
2 :  [0 1 0 0 0 0 0 0 0 0] : POP060210
7 :  [0 0 0 0 0 0 1 0 0 0] : RHI125214
26 :  [0 1 0 0 0 0 1 0 0 0] : POP060210 RHI125214
28 :  [0 1 0 0 0 0 0 0 1 0] : POP060210 HSG495213
46 :  [0 0 0 0 1 1 0 0 0 0] : SEX255214 SBO515207
55 :  [0 0 0 0 0 1 0 0 0 1] : SBO515207 PVY020213
58 :  [0 0 0 0 0 0 1 0 1 0] : RHI125214 HSG495213
Train
Accuracy: 0.8309031044214488
F1: 0.8222496909765142
auc:  0.9136786053422706
Test
Accuracy: 0.8426073131955485
F1: 0.5991902834008097
auc:  0.8484660966096609
```

Cluster 1

```
cluster: 1  indices: [0, 1, 3, 21, 23, 27, 32, 33, 35, 36, 43, 45, 47, 48, 50]
('PST045214', 'Population, 2014 estimate')
('PST040210', 'Population, 2010 (April 1) estimates base')
('POP010210', 'Population, 2010')
('VET605213', 'Veterans, 2009-2013')
('HSG010214', 'Housing units, 2014')
('HSD410213', 'Households, 2009-2013')
('BZA010213', 'Private nonfarm establishments, 2013')
('BZA110213', 'Private nonfarm employment,  2013')
('NES010213', 'Nonemployer establishments, 2013')
('SBO001207', 'Total number of firms, 2007')
('MAN450207', 'Manufacturers shipments, 2007 ($1,000)')
('RTN130207', 'Retail sales, 2007 ($1,000)')
('AFN120207', 'Accommodation and food services sales, 2007 ($1,000)')
('BPS030214', 'Building permits, 2014')
('POP060210', 'Population per square mile, 2010')
```

Cluster 2

```
cluster: 2  indices: [2, 6, 11, 25, 37, 38, 39, 41, 42, 44, 46]
('PST120214', 'Population, percent change - April 1, 2010 to July 1, 2014')
('AGE775214', 'Persons 65 years and over, percent, 2014')
('RHI425214', 'Asian alone, percent, 2014')
('HSG096213', 'Housing units in multi-unit structures, percent, 2009-2013')
('SBO315207', 'Black-owned firms, percent, 2007')
('SBO115207', 'American Indian- and Alaska Native-owned firms, percent, 2007')
('SBO215207', 'Asian-owned firms, percent, 2007')
('SBO415207', 'Hispanic-owned firms, percent, 2007')
('SBO015207', 'Women-owned firms, percent, 2007')
('WTN220207', 'Merchant wholesaler sales, 2007 ($1,000)')
('RTN131207', 'Retail sales per capita, 2007')
```

Cluster 3

```
cluster: 3  indices: [19, 20, 26, 29, 30, 31]
('EDU635213', 'High school graduate or higher, percent of persons age 25+, 2009-2013')
('EDU685213', "Bachelor's degree or higher, percent of persons age 25+, 2009-2013")
('HSG495213', 'Median value of owner-occupied housing units, 2009-2013')
('INC910213', 'Per capita money income in past 12 months (2013 dollars), 2009-2013')
('INC110213', 'Median household income, 2009-2013')
('PVY020213', 'Persons below poverty level, percent, 2009-2013')
```

Cluster 4

```
cluster: 4  indices: [8, 9, 12, 13, 14, 15, 16, 17, 18, 24]
('RHI125214', 'White alone, percent, 2014')
('RHI225214', 'Black or African American alone, percent, 2014')
('RHI525214', 'Native Hawaiian and Other Pacific Islander alone, percent, 2014')
('RHI625214', 'Two or More Races, percent, 2014')
('RHI725214', 'Hispanic or Latino, percent, 2014')
('RHI825214', 'White alone, not Hispanic or Latino, percent, 2014')
('POP715213', 'Living in same house 1 year & over, percent, 2009-2013')
('POP645213', 'Foreign born persons, percent, 2009-2013')
('POP815213', 'Language other than English spoken at home, pct age 5+, 2009-2013')
('HSG445213', 'Homeownership rate, 2009-2013')
```

Cluster 5

```
cluster: 5  indices: [4, 5, 28]
('AGE135214', 'Persons under 5 years, percent, 2014')
('AGE295214', 'Persons under 18 years, percent, 2014')
('HSD310213', 'Persons per household, 2009-2013')
```

Cluster 6

```
cluster: 6  indices: [7, 10, 22, 34, 40, 49]
('SEX255214', 'Female persons, percent, 2014')
('RHI325214', 'American Indian and Alaska Native alone, percent, 2014')
('LFE305213', 'Mean travel time to work (minutes), workers age 16+, 2009-2013')
('BZA115213', 'Private nonfarm employment, percent change, 2012-2013')
('SBO515207', 'Native Hawaiian- and Other Pacific Islander-owned firms, percent, 2007')
('LND110210', 'Land area in square miles, 2010')
```