Demographic Predictors of Election Results

Cindy Li & Maya Dayan

Background

Background

Most State Polls Were Not Weighted by Education

	Weighted by education	Unknown	Did not weight by education
National polls	49%	28	23
Battleground states	32	20	48
Battleground, non-online	14	19	67
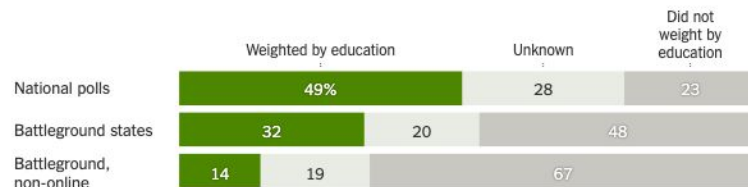
The postelection shift is also potentially consistent with a subtly different explanation: the so-called "shy Trump" effect. The idea is that Trump supporters took telephone surveys but were embarrassed to divulge their support for an unpopular candidate. If true, the "undecided" voters were really Trump voters all along; they just didn't want to admit it to pollsters until after their candidate won.

Among validated voters in 2016, wide gap among whites by education

% of validated voters in 2016 who reported voting for …

	Clinton	Trump	Share of electorate
Total	48	45	
Men	41	52	45
Women	54	39	55
18-29	58	28	13
30-49	51	40	30
50-64	45	51	29
65+	44	53	27
White	39	54	74
Black	91	6	10
Hispanic	66	28	10
White men	32	62	33
White women	45	47	41
Black men	81	14	4
Black women	98	*	6
Hispanic men	65	28	5
Hispanic women	67	28	5
College grad +	57	36	37
Non-college grad	43	50	63
Among whites College grad +	55	38	30
Non-college grad	28	64	44
Among nonwhites College grad +	68	26	7
Non-college grad	77	18	19

Research Question

What demographic indicators are the best predictors of a county's voting patterns, both in general and primary elections? How strong are purely demographic prediction models (without polling data)?

Data

- Census demographics (2014, 2016, 2018):
    - *https://www.census.gov/data/developers/data-sets/acs-5year.html*
- Presidential election outcomes (2016):
    - *https://raw.githubusercontent.com/tonmcg/US_County_Level_Election_Results_08-16/master/2016_US_County_Level_Presidential_Results.csv*
- Governor's race data (2014):
    - *https://int.nyt.com/applications/elections/2014/data/2014-11-04/supermap/governor.json*
- Primary results (2020):
    - https://int.nyt.com/applications/elections/2020/data/api/2020-02-03/iowa/president/democrat.json
    - https://int.nyt.com/applications/elections/2020/data/api/2020-02-11/new-hampshire/president/democrat.json
    - https://int.nyt.com/applications/elections/2020/data/api/2020-02-22/nevada/president/democrat.json
    - https://int.nyt.com/applications/elections/2020/data/api/2020-02-29/south-carolina/president/democrat.json
    - https://int.nyt.com/applications/elections/2020/data/api/2020-03-03/california.json
    - https://int.nyt.com/applications/elections/2020/data/api/2020-03-03/texas.json

Model

We trained several models:
- Trees: Decision Tree, Random Forest, XGBoosted Tree
- Logistic Regression, SVM (radial)
- K-Nearest Neighbor

We trained our models based on demographic data separately for two tasks:
- predict Democratic primary election results
- predict whether a county would vote Democrat or Republican in the 2016 presidential election
  - trained on 2014 governor election results

Results

Presidential Election 2016

training set: IA, MI, PA, WI - 2014 demographics & governor election results

testing set: IA, MI, PA, WI - 2016 demographics

Decision Tree:
- Train accuracy: 0.9345794
- Test accuracy: 0.7912773

Random Forest
- Train accuracy: 1
- Test accuracy: 0.8971963

SVM (radial)
- Train accuracy: 0.8380062
- Test accuracy: 0.8816199

Logistic Regression
- Train accuracy: 0.8660436
- Test accuracy: 0.8660436

KNN
- Train accuracy: 0.8411215
- Test accuracy: 0.8660436

Decision Tree

1 indicates Republican
0 indicates Democrat
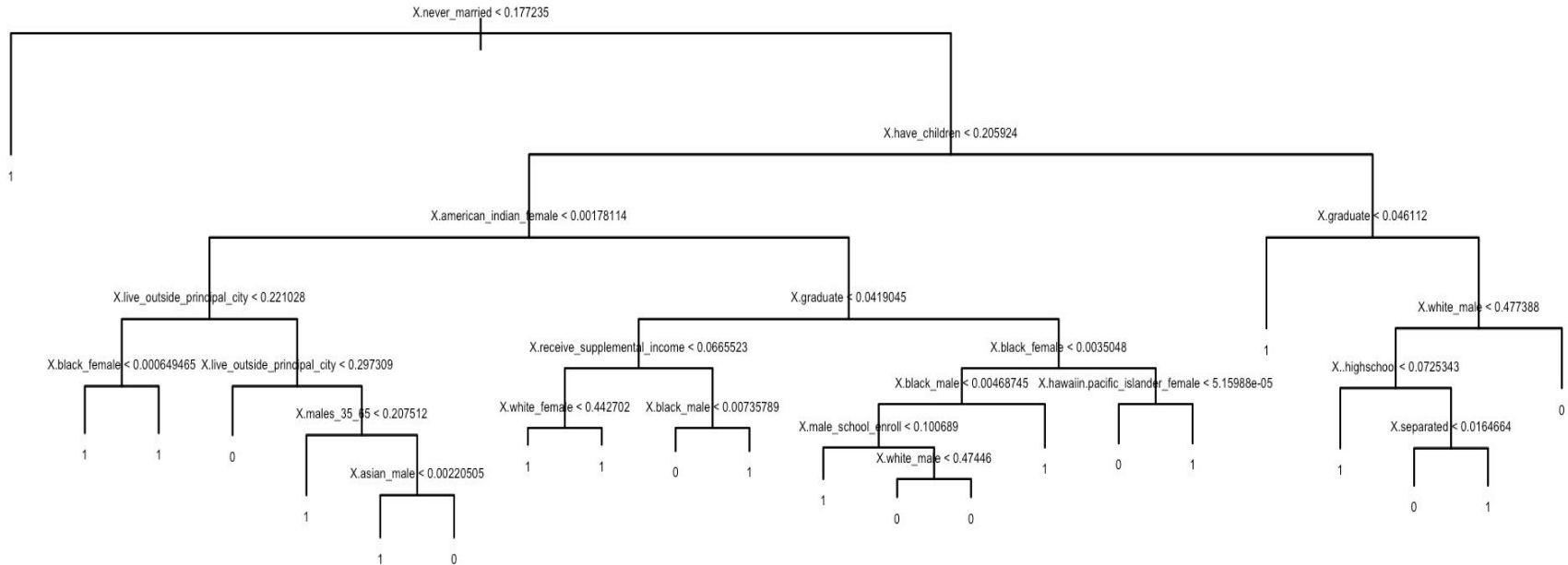
Feature Importance *(from random forest)*
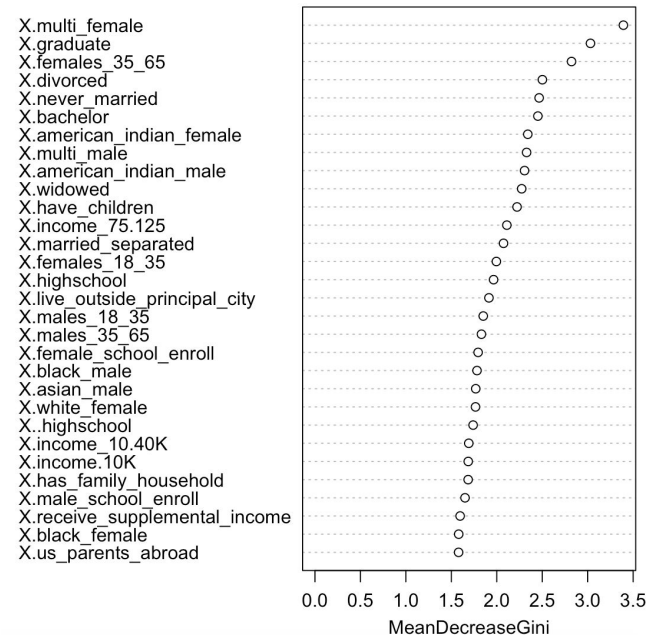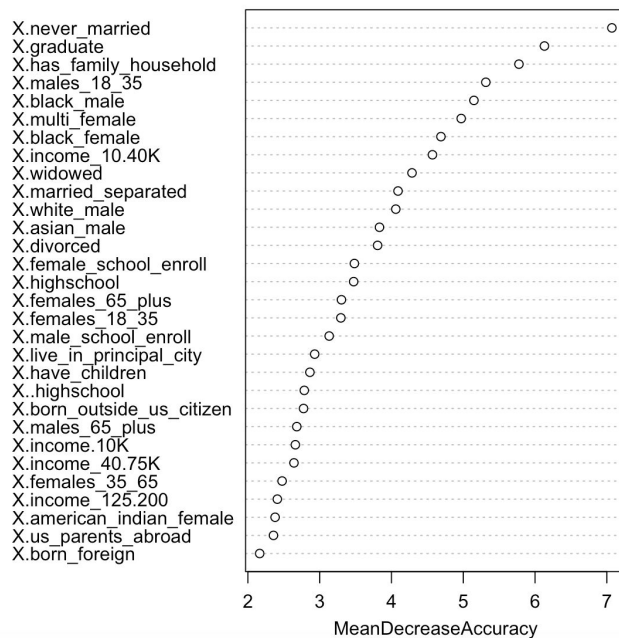
note on feature names: some symbols in the feature names were lost when loading data. for example, X..highschool is supposed to indicate an education level < highschool. likewise, X.income.10K indicates income < $10,000 while income X.income_125.200 indicates income between $125,000 and $200,000

Feature Importance *(from logistic regression)*

```
Call:
glm(formula = as.factor(winner) ~ ., family = binomial, data = train)

Deviance Residuals:
    Min      1Q   Median      3Q     Max
-2.5932  0.0374   0.2403  0.5245  1.8104

Coefficients: (4 not defined because of singularities)
                                    Estimate Std. Error z value Pr(>|z|)
(Intercept)                         -137.846    129.815  -1.062  0.28830
X.white_male                         411.360    156.311   2.632  0.00850 **
X.white_female                        85.140    116.404   0.731  0.46452
X.black_male                         395.048    154.501   2.557  0.01056 *
X.black_female                       123.656    113.230   1.092  0.27480
X.american_indian_male               218.474    163.602   1.335  0.18175
X.american_indian_female             251.465    135.662   1.854  0.06379 .
X.asian_male                         364.844    182.606   1.998  0.04572 *
X.asian_female                       233.526    162.213   1.440  0.14997
X.hawaiin.pacific_islander_male       71.361    446.591   0.160  0.87305
X.hawaiin.pacific_islander_female   -301.008    810.464  -0.371  0.71034
X.other_male                         512.649    224.553   2.283  0.02243 *
X.other_female                         7.863    226.344   0.035  0.97229
X.multi_male                         579.110    210.540   2.751  0.00595 **
X.multi_female                            NA         NA      NA       NA
X.latino_male                       -124.896     91.286  -1.368  0.17125
X.latino_female                      127.952     95.294   1.343  0.17937
X.us_born                            -59.474     41.802  -1.423  0.15481
X.us_island_born                    -115.626     74.494  -1.552  0.12062
X.us_parents_abroad                  161.897    157.043   1.031  0.30258
X.naturalized                        -95.197     69.111  -1.377  0.16838
X.not_citizen                             NA         NA      NA       NA
X.born_outside_us_citizen                 NA         NA      NA       NA
X.born_foreign                            NA         NA      NA       NA
X.never_married                     -115.503    128.700  -0.897  0.36947
X.married_separated                  -85.106    130.304  -0.653  0.51367
X.divorced                           -94.243    132.059  -0.714  0.47545
X.separated                          -55.153    142.720  -0.386  0.69917
X.widowed                             61.911    136.705   0.453  0.65064
X..highschool                         -2.846     55.768  -0.051  0.95931
X.highschool                          -3.069     49.062  -0.063  0.95012
X..college                           -11.280     46.840  -0.241  0.80969
X.bachelor                             7.773     52.384   0.148  0.88204
X.graduate                           -59.309     52.431  -1.131  0.25798
X.live_in_principal_city              -4.366      3.332  -1.310  0.19004
X.live_outside_principal_city         -4.308      1.671  -2.579  0.00991 **
X.have_children                      -59.119     59.638  -0.991  0.32154
X.receive_supplemental_income         -8.607     29.236  -0.294  0.76847
X.has_family_household                69.735     50.108   1.392  0.16401
X.male_school_enroll                 -26.854     48.803  -0.550  0.58215
X.female_school_enroll                61.756     51.663   1.195  0.23195
X.income.10K                          55.626     50.668   1.098  0.27227
X.income.200K                         52.048    101.676   0.512  0.60872
X.males_18_35                        -79.345    134.378  -0.590  0.55488
X.males_35_65                        -92.106    141.604  -0.650  0.51540
X.males_65_plus                     -122.743    150.660  -0.815  0.41524
X.females_18_35                      156.812    162.459   0.965  0.33442
X.females_35_65                       88.041    189.808   0.464  0.64276
X.females_65_plus                    172.513    192.342   0.897  0.36977
X.income_10.40K                      -57.350     30.023  -1.910  0.05611 .
X.income_40.75K                       31.376     37.317   0.841  0.40046
X.income_75.125                       44.590     42.794   1.042  0.29743
X.income_125.200                      -4.331     68.922  -0.063  0.94989
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 290.86  on 320  degrees of freedom
Residual deviance: 194.30  on 272  degrees of freedom
AIC: 292.3

Number of Fisher Scoring iterations: 7
```
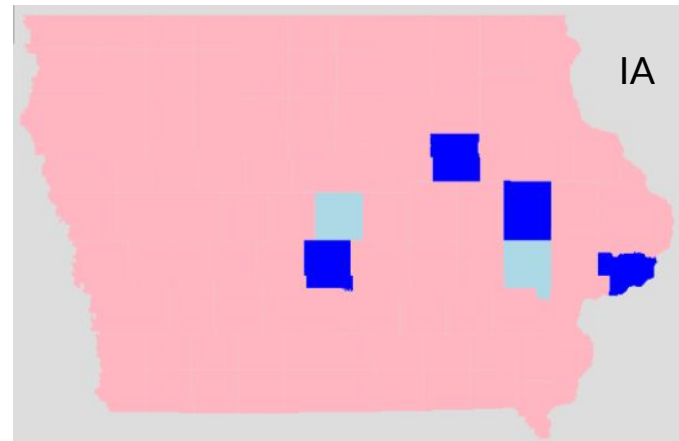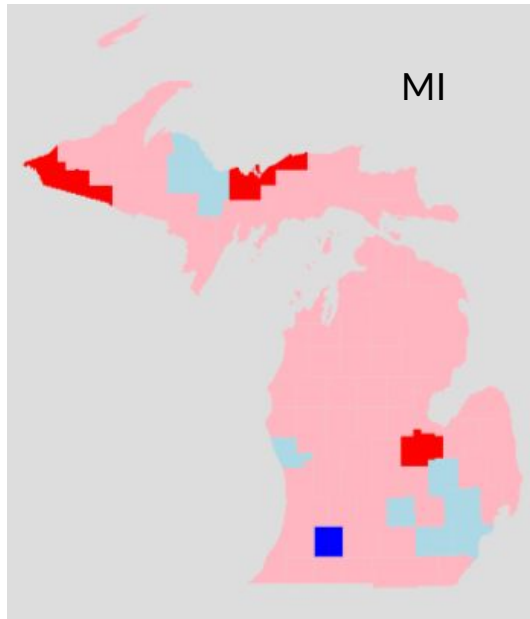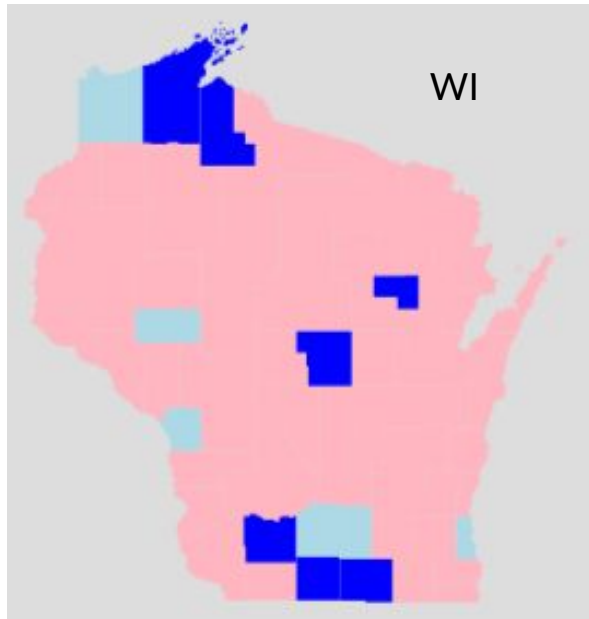
Accuracy

Bright Red indicates that Republicans won but we predicted Democrats would win

Bright Blue indicates that Democrats won but we predicted Republicans would win

Light Colors indicate our prediction was correct (R in Red, D in Blue)

WI

MI

IA

PA

Primaries 2020

training set: IA, NH, NV, SC primary results

testing set: CA, TX primary results

Decision Tree:
- Train accuracy: 0.8546512
- Test accuracy: 0.5544872

Random Forest
- Train accuracy: 1
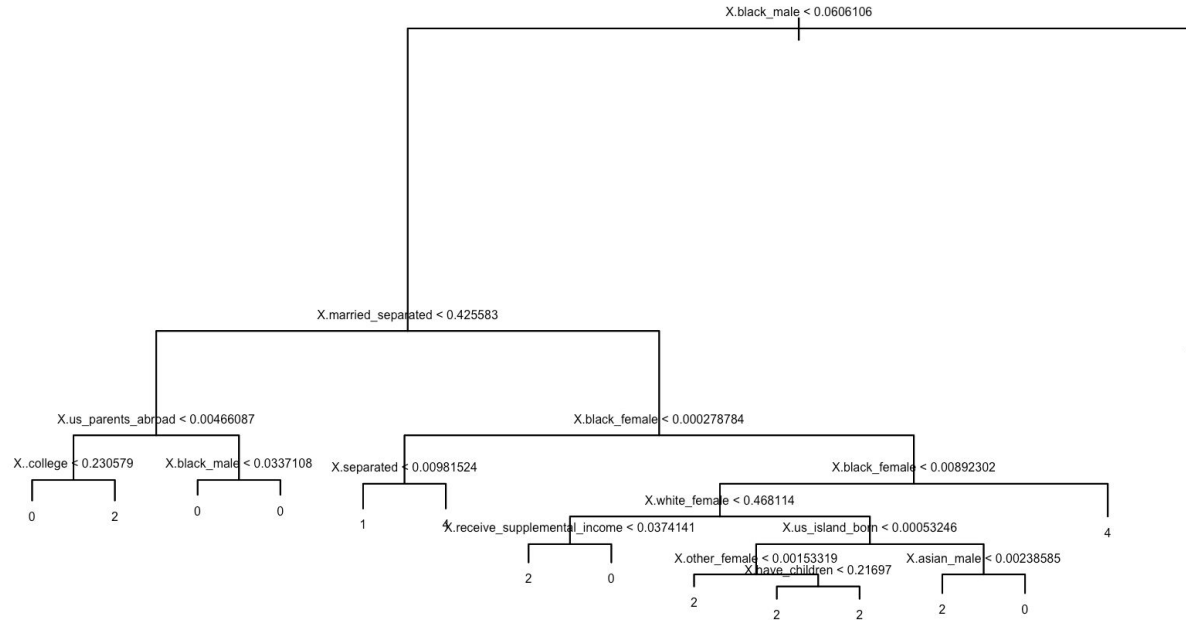- Test accuracy: 0.4839744

XGBoost
- Train accuracy: 0.8604651
- Test accuracy: 0.4455128

KNN
- Train accuracy: 0.7906977
- Test accuracy: 0.3717949

Decision Tree

0 indicates Sanders   2 indicates Buttigieg
1 indicates Klobuchar   4 indicates Biden

Conclusion

- The demographics predicted whether a county voted Republican or Democratic in 2016 with relatively high accuracy using a model trained on 2014 data
- If this model holds true, it can be useful predicting the outcomes in swing states
- Demographics were less indicative of primary voting