OR/SYST 568 Term Project Proposal

Project Name: 2016 Presidential Election Predictions

Team Members:  
 Anthony Calisti, David Croghan, Alec Morris, Brad Wilkins, Kevin Conklin, & Preet Pal Singh

* Prediction Problem

Data analytics have played a large part in presidential elections since President Obama’s successful re-election campaign in 2012. (Issenberg, 2012). Campaign workers analyze data to target undecided regions and ignore those areas that have made up their mind (for or against). Group 4’s proposed project reviews voting data from the 2008, 2012, and 2016 presidential campaigns. The primary objective is to use information from the different campaigns to predict the election outcomes of 2016 for each county recorded. Some questions that may be answered include:

* Which predictors were most valuable when predicting who a county would vote for?
  + Is this outcome surprising or expected given political trends and information?
* How might 3rd party voting have swung the outcome?
  + Did the Green Party take voters from the Democrats?
  + Did libertarian/write in/other conservatives take voters from the Republicans?
* Would this model perform similarly in 2020?

Each record in the dataset represents a state, one of its counties, and 162 attributes about the county including demographics and level of education. Many of these attributes will be useful as predictors while others can be removed from the model. For example, “Geo Shape” defines a county’s border using geospatial coordinates and would not be considered categorical or continuous. Additional details follow in “Data Preprocessing.”

Election results represent the response variable. The 2016 election results will be used to verify the accuracy of the model (Test Set). To minimize the chance of overfitting the model, techniques like 10-fold cross validation will be used to break up the data into training and validation sets.

* Dataset & Preprocessing

The dataset consists of 164 data fields covering 3145 counties including the number of voters, the percentage of the electorate voting for each party in 2008, 2012 and 2016, demographics, education and health statistics, economic data, as well as information about county location, climate and population and more. The dependent variable is binary, showing whether or not the Republican Party won the 2016 election for a given county. While categorical and binary variables are represented in the predictor variables, the majority are continuous as they are often represented as percentages of the population (percentage Hispanic, percentage with a high school diploma, etc.)

Given the size of the dataset, the first step in our data preprocessing would be factor selection and reduction. From an initial count of null values, 43 predictor variables have null values for more than 1400 rows. These values should likely be removed. Any of the categorical and binary variables that have very low variability should also be removed. The dataset has many predictors that intuitively would be related. For example, the rate of obesity and the rate of diabetes in a county are both provided variables. Using correlation analysis we would eliminate predictors that are highly correlated with other predictor variables. In the example above, if diabetes and obesity are related, we would remove the one that has the highest average correlation with the other predictor variables. Depending on the model we choose, we may want to further narrow down the number of variables we use, both to avoid over-fitting the model, and to maintain interpretability. This would be especially important in a linear regression model. One way to limit the number of predictor variables would be to use PCA. While this would certainly limit the number of predictors, it tends to weaken interpretability. Another option would be to use PCA initially to view the coefficients on each of the predictor variables in the first couple of components to get a sense of which variables capture the most variability. Given the number of predictor variables, there is a lot of room for factor engineering by using variable interactions, or changing the way variables are presented (for example, total number of college graduates versus percentage).

The next step of the preprocessing would be to normalize distributions with heavy skew, high kurtosis or bimodality. Because most of the continuous variables or population percentages, there should not be any negatives. In this case Box Cox transformation could be used to optimally normalize each of the variables.

The last step would be to deal with missing values. While we eliminated the columns with half of their rows missing, there are still a large number of columns with missing data. We could use imputation to fill in these variables, where the method of imputation would depend on the model. For simpler models like a linear regression, it might be best to use the median to prevent loss of interpretability.

* Measure the Quality of Predictions
  + We can easily measure the quality of our predictions by comparing to the actual 2016 election results by county

Ultimately all the analysis boils down to a simple binary prediction of which candidate a county voted for. Since all counties will have a prediction and all predictions will be for one of two candidates, a simple percentage accuracy should suffice. Other projects with a binary output might be better served by a scoring system which weights false positive and true negative (for example, when identifying potential suicide prone individuals, it is likely better to have more false positives) but in this case there is no particular weight.

Data Columns (3143 rows)

|  |  |  |  |
| --- | --- | --- | --- |
| State | Sales.and.office.occupations | annual\_PRCP | votes16\_kahnl |
| ST | Farming.fishing.and.forestry.occupations | winter\_PRCP | votes16\_la\_rivag |
| Fips | Construction.extraction.maintenance.and.repair.occupations | summer\_PRCP | votes16\_hoeflingt |
| County | Production.transportation.and.material.moving.occupations | spring\_PRCP | votes16\_kenistonc |
| precincts | White | autumn\_PRCP | votes16\_smithm |
| votes | Black | annual\_TAVG | votes16\_atwoodf |
| Democrats 08 | Hispanic | annual\_TMAX | votes16\_kennedya |
| Democrats 12 | Asian | annual\_TMIN | votes16\_kopitkek |
| Republicans 08 | Amerindian | winter\_TAVG | votes16\_kotlikoffl |
| Republicans 12 | Other | winter\_TMAX | votes16\_lyttleb |
| Republicans 2016 | White\_Asian | winter\_TMIN | votes16\_maldonadoj |
| Democrats 2016 | SIRE\_homogeneity | summer\_TAVG | votes16\_maturenm |
| Green 2016 | median\_age | summer\_TMAX | votes16\_scottr |
| Libertarians 2016 | lon | summer\_TMIN | votes16\_silvar |
| Republicans 2012 | lat | spring\_TAVG | votes16\_soltysike |
| Republicans 2008 | Poor.physical.health.days | spring\_TMAX | votes16\_vacekd |
| Democrats 2012 | Poor.mental.health.days | spring\_TMIN | votes16\_copelands |
| Democrats 2008 | Low.birthweight | autumn\_TAVG | votes16\_jacobp |
| X1 | Teen.births | autumn\_TMAX | votes16\_whitej |
| Less.Than.High.School | Children.in.single.parent.households | autumn\_TMIN | votes16\_mooreheadm |
| At.Least.High.School.Diploma | Adult.smoking | nearest\_county | votes16\_none\_of\_these\_candidates |
| At.Least.Bachelor.s.Degree | Adult.obesity | temp | votes16\_duncanr |
| Graduate.Degree | Diabetes | precip | votes16\_skewesp |
| School.Enrollment | Sexually.transmitted.infections | temp\_bins | votes16\_giordanir |
| Median.Earnings.2010.dollars | HIV.prevalence.rate | lat\_bins | total16 |
| White.not.Latino.Population | Uninsured | lon\_bins | other16\_frac |
| African.American.Population | Unemployment | precip\_bins | rep16\_frac2 |
| Native.American.Population | Violent.crime | elevation\_bins | dem16\_frac2 |
| Asian.American.Population | Homicide.rate | Geo Shape | name\_prev |
| Population.some.other.race.or.races | Injury.deaths | name\_16 | statecode\_prev |
| Latino.Population | Infant.mortality | reporting | total08 |
| Children.Under.6.Living.in.Poverty | CA | votes16\_trumpd | total12 |
| Adults.65.and.Older.Living.in.Poverty | S | votes16\_clintonh | other08 |
| Total.Population | MAR | votes16\_johnsong | other12 |
| Preschool.Enrollment.Ratio.enrolled.ages.3.and.4 | CFS | votes16\_steinj | other12\_frac |
| Poverty.Rate.below.federal.poverty.threshold | ACFS | votes16\_castled | other08\_frac |
| Gini.Coefficient | MeanALC | votes16\_de\_la\_fuenter | rep12\_frac2 |
| Child.Poverty.living.in.families.below.the.poverty.line | MaxALC | est\_votes\_remaining | rep08\_frac2 |
| Management.professional.and.related.occupations | Mixedness | votes16\_mcmulline | dem12\_frac2 |
| Service.occupations | elevation | votes16\_hedgesj | dem08\_frac2 |

# Works Cited

Issenberg, S. (2012, December 19). *How Obama's Team Used Big Data to Rally Voters.* Retrieved from MIT Technology Review: https://www.technologyreview.com/s/509026/how-obamas-team-used-big-data-to-rally-voters/