Graves DilEmma: Anna Graves,
Joshua Graves,
Emma Johnson

Data Science Team Project #2 Write-Up

Generally speaking, there are many characteristics that play a role in determining voting behavior. Age, geographic location, highest educational attainment, income, and religious beliefs all likely influence what political party someone thinks aligns most closely with their interests. Therefore, each of the variables provided have a reasonable political argument for being used to predict voting outcomes. We especially wanted variables that provided information on population and race. Population is a key variable to analyze because coastal states and areas with high populations tend to vote Democrat. This includes big cities like New York, Chicago, and Los Angeles, which all went to Hillary Clinton by large margins. There also tend to be differences in voting behavior along racial lines. For example, a majority of Black Americans voted for Hillary Clinton. Additionally, we performed a "relative spread" check and a correlation check to provide mathematical support for the variables we selected.

For our analysis, we use the variables: WTN220207 (renamed as Merch_sales), PST045214 (renamed as Pop), RHI825214 (renamed as Perc_white), RHI225214 (renamed as Perc_black), EDU685213 (renamed as Perc_bachelors_plus), HSG096213 (renamed as Perc_housing), and POP815213 (renamed as Perc_no_eng). In order to calculate relative spread, we first looked for variables that had the widest average spread based on political party. We define "spread" as the difference between democrat and republican votes. For example, the average Hispanic or Latino percent (2014) for counties that voted Democrat was 16.15%, whereas the counties that voted Republican had an average Hispanic or Latino percent of 7.73%. Therefore, the spread is 16.15%-7.73%, which equals 8.42%. This seems to indicate that areas with a higher Hispanic or Latino percentage vote more for Democrats. While the averages alone might be useful for the model, spread can show if there is a large natural distinction between the averages by party. However, some of the variables are based on percentages, while others are in dollar amounts. Thus, using the "relative spread" takes into account the respective proportions, which in turn makes it easier to find the largest relative difference. We then calculated the

correlation between similar variables (i.e. population and number of housing units) to avoid multicollinearity using a correlation matrix. Any variable with a correlation above 0.65 was dropped, with the exception of Population. The table below provides a brief explanation of each variable and a short explanation for how we ended up with the seven variables used. Although this method is helpful in choosing variables to include, one possible downside is the risk of overfitting the model because the averages in this sample may draw distinctions in the spread that do not exist on a national level. Even so, a sample should approximate the true national averages so perhaps this methodology retains some validity.

Name	New Name	Description	Dem	Rep	Spread	Kept	Justification
WTN220207	**Merch_sales**	Merchant wholesaler sales, 2007 ($)	6,100,499.00	385,902.10	5,714,596.90	Yes	Good indicator towards dem.
MAN450207	N/A	Manufacturers shipments, 2007 ($)	5,305,693.00	822,709.00	4,482,984.00	No	High correlation with Merchant variable
RTN130207	N/A	Retail sales, 2007 ($1,000)	4,916,453.90	646,712.60	4,269,741.30	No	See above
AFN120207	N/A	Accommodation and food services sales, 2007 ($1,000)	900,472.51	77,825.24	822,647.27	No	See above
PST045214	**Pop**	Population, 2014 estimate	387,133.97	55,588.44	331,545.53	Yes	Population size is good indicator
PST040210	N/A	Population, 2010 (April 1) estimates base	372,946.07	54,209.27	318,736.80	No	Population 2014 was better indicator
POP010210	N/A	Population, 2010	372,933.30	54,204.41	318,728.89	No	See above
BZA110213	N/A	Private nonfarm employment, 2013	157,290.17	16,981.02	140,309.15	No	High correlation with multiple variables
HSG010214	N/A	Housing units, 2014	155,904.54	24,494.41	131,410.13	No	Housing unit percent was better
HSD410213	N/A	Households, 2009-2013	137,638.58	20,627.82	117,010.76	No	See above
VET605213	N/A	Veterans, 2009-2013	21,460.75	4,430.02	17,030.72	No	High correlation with population
BZA010213	N/A	Private nonfarm establishments, 2013	9,501.33	1,205.84	8,295.50	No	High correlation with Merchant.
RHI825214	**Perc_white**	White alone, not Hispanic or Latino, percent, 2014	55.36	81.68	26.32	Yes	Good spread, important
RHI125214	N/A	White alone, percent, 2014	69.62	88.56	18.93	No	White alone, not Hispanic was better
RHI225214	**Perc_black**	Black or African American alone, percent, 2014	21.47	6.91	14.56	Yes	Good spread, important
HSG096213	**Perc_housing**	Housing units in multi-unit structures, percent, 2009-2013	23.34	10.42	12.92	Yes	See above
POP815213	**Perc_no_eng**	Language other than English spoken at home, 2009-2013	18.13	7.46	10.68	Yes	See above
EDU685213	**Perc_bachelors_plus**	Bachelor's degree or higher, % persons 25+, 2009-2013	28.09	18.20	9.89	Yes	See above, education important

After selecting our seven variables, we ran a series of models in order to compare the accuracy, precision, and recall of each and decide on the best one. First, we fit a logistic

regression model and estimated the cross-validation error. Using a threshold of 0.5, the accuracy was 91.9 percent, the precision was 93.6 percent, and the recall was 97.1 percent. The cross-validation error rate was 8.1 percent. Using a threshold of 0.05, recall increases to 99.9 percent, while both accuracy and precision drop to 87.6 percent and 87.3 percent respectively. This cross-validation error rate was 12.4 percent. Although we found that decreases in accuracy and precision led to an increase in recall, we decided to hold the threshold at 0.5; a necessary tradeoff in the quest of a good-fitting model.

Next we ran two versions of the k-nearest neighbors algorithm to see if our results improved. We split our data 80/20 into a training and testing group, respectively. The first version uses all seven variables. This results in an accuracy of 0.87, a precision of 0.88, and a recall of 0.98. The second version only uses the variables Perc_White, Per_Bachlors_Plus, and Perc_Black. Despite using fewer variables, this results in a higher accuracy of 0.92, a higher precision of 0.94, and a higher recall of 0.97. We also ran a LOOCV for the train data before predicting the accuracy, precision, and recall of each version in hopes of minimizing the test error.

Next we decided to try a random forest model to see if we could find a better fitted model than k-nearest neighbors. We fit a model using our 7 variables and set ntree=500 and mtry=3. This model yielded the best results with an accuracy of 93.1 percent, a precision of 94.8 percent, and a recall of 97.1 percent. Additionally, we did not cross validate our random forest model because our OOB model is essentially doing the same thing as cross-validation. Considering that the accuracy, precision, and recall were the highest we chose the random forest classification model to predict pred_winner in the test data set.

Additional Information:

Florida, Richard. *How America's Metro Areas Voted*. November 29, 2016.
https://www.citylab.com/equity/2016/11/how-americas-metro-areas-voted/508355/