Milestone Report

Pramod Padmanabhan

July 21, 2016

The Problem

This project is about the United States election, 2016. We will be studying the charactersitics of the voters of Donald Trump, who to the surprise of the entire world has almost won the nomination of the Republican party to run for the president's election to be held in November, 2016. We would like to find out what percentage of the Republican voters from each county of every state vote for Trump and study their characteristics classified by their race, education, if they are in urban areas, what their professions are, their levels of employment, if they are in poverty stricken areas, their mean household income, and if they have extremist views, in particular if they belong to extremist groups such as the Ku Klux Klan (KKK).

The problem thus invlolves collecting data sets from different sources to create one big data set that contain all the above mentioned variables. This involves a reasonable amount of data wrangling. This component of the project cover two important aspects of a data science problem - relating seemingly unrelated variables to predict an outcome and cleaning data sets to bring it to a form where we can work with exactly these features.

Having obtained the data set the problem provides us ample scope to analyse it using linear regression and regression trees to predict the votes for Donald Trump. We do not expect a perfect linear correlation between our dependent and independent variables and hence we will need to resort to clustering and regression trees to get around this.

The Data Sets and their Cleaning

To study Trump's voters we require different data sets which meet our requirements.

The Dependent Variable

The data set which contains the dependent variable, namely the votes gained by Trump can be found in https://www.kaggle.com/benhamner/2016-us-election. This data set includes the results from the primaries of 46 states held for the Republican candidates. It does not include Colorado, DC, Maine, Minnesota and North Dakota. The primaries are held in each county of each state and so the data is the number of votes for each of the counties. However we will only use values for each state and so we group by the state variable and sum the votes for each county. Note that these votes are only from the registered voters who have declared themselves as Republicans and this explains the low numbers in this column.

The data also includes the primary results for the Democrats where there are 49 observations. We will only use this partially to make a preliminary analysis on the characteristics of the voters of Hillary Clinton to make a comparison with our test subject, Donald Turmp.

From this data set we filter out the rows containing "Donald Trump" from the candidate variable. This gives us the dependent variable for the 46 states.

The Predictors

We use 18 variabels to study our dependent variable. These are grouped together according to correlations amongst them.

We have the **Hate group** variable which describes the number of Ku Klux Klans in all the counties of all states. This data set can be found at http://scholarscompass.vcu.edu/hist_data/1/. As the data is for each county we again group by the state variable and compute the number of Klans for each state. And now since the states vary in size, the number of counties in each state also vary which forces us to normalize our observation by the county number. We will be carrying out this procedure for each of the predictors to be listed below and hence will not be repeating this.

Next we have the **Education/Employment/Poverty Variables** which give us information about when counties are considered as having a low amount of education, low employment and when it is considered as having persistent poverty. This data was taken from http://www.ers.usda.gov/data-products/county-typology-codes.aspx. These variables were again normalised according to the procedure described previously. These three variables are bound to have correlations and so we group them together and study the dependent variable with respect to them as a graoup.

We then have the **Economic Variables** which describe the dominant occupation of each county in every state. The occupations considered are **Farming**, **Mining**, **Manufacturing**, **Federal/State Government** jobs, **Recreation** fields and finally **Nonspecialized** counties which do not indulge in any of the latter professions. These variables are again normalised and used to predict the percentage of votes recieved by Trump. These variables are mostly independent and could have a bit of correlations. We group them together as economic variables.

We then consider the **Urbanization** variable which measures the number of metors in each county according to a fixed definition. They indicate howw urban a given state is and is taken as a single predictor after being normalised as before.

The next group of variables belong to the different **Races**, where the percentages of the different racial groups in each county of every state are given. The racial groups considered are the **Whites**, **Blacks**, **Hispanics**, **Asians**, **Natives**, **Pacific Islanders** and those of **Mixed** origin. We again normalise each of these variables before using them for analysis.

Finally we have the **Income** variable which provides statistics for the mean household income of each county in every state. This is again normalised before use.

There are several other variables which we could have used to study the characteristics of Trump's voters. Among these the most useful is the Religion variable where we have information about the religious composition of each state. We did find the data set in http://www.thearda.com/rcms2010/ but we could only obtain information about the various Christian groups in each county. This was nevertheless added to the final data set but we did not find it useful as there was not enough equivalent information about other religions. Hence we dropped this 'crucial' variable.

We also had information about the distribution of Veterans in every state, obtained from http://www.va.gov/vetdata/Veteran_Population.asp. We did not use this as we felt it may not be that significant for our analysis.

The data cleaning was pretty much using the filter and summarise functions from dplyr and using some functions to standardise the names of the observations for the different variables. Most of our data, apart from the names of the states is just numbers. However there were many hurdles in joining the data sets together as they were not all of the same class which was taken care of before combining them. Most of the variables were of class factor which had to be converted to numeric before applying any arithmetic function on them.

Preliminary Analysis

We first a few basic scatter plots between the dependent variable and the predictors and attempted a linear regression model on the resulting plot. We found that the correlation was very low as it had a low R-squared value.

This can be seen for example in the data between Votes and Klan percentage for every state.

```
##
## Call:
## lm(formula = absolute_votes ~ Klans, data = klan)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.2444  -3.6350   0.2736   3.3910  17.8041
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.116e+01  1.058e+00  10.547 1.26e-13 ***
## Klans       9.685e-05  1.163e-04   0.833    0.409
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.252 on 44 degrees of freedom
## Multiple R-squared:  0.01552,    Adjusted R-squared:  -0.00685
## F-statistic: 0.6939 on 1 and 44 DF,  p-value: 0.4094
```

This shows a low R-squared value and so we resort to regression trees.

Similary for the dependence on education, employment and poverty variables we can find the R-squared values

```
##
## Call:
## lm(formula = Votes ~ Education + Employment + Poverty, data = EEP1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.6501  -3.6342  -0.0958   3.0731  18.4262
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 10.65011    1.26896   8.393 1.59e-10 ***
## Education   -0.04135    0.09644  -0.429   0.6703
## Employment   0.08519    0.03802   2.241   0.0304 *
## Poverty     -0.10490    0.08816  -1.190   0.2407
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.089 on 42 degrees of freedom
## Multiple R-squared:  0.1088, Adjusted R-squared:  0.04514
## F-statistic: 1.709 on 3 and 42 DF,  p-value: 0.1797
```

Again due to the low R-squared value we turn to regression trees.

For economic indicators we have

```
##
## Call:
## lm(formula = Votes ~ Farming + Mining + Manufacturing + Government +
##     Recreation + Nonspecialized, data = economy1)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.3074  -3.4493  -0.2794   4.1226  16.7988
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)    11.5504905  1.7640379   6.548    9e-08 ***
## Farming        -0.0094972  0.0136154  -0.698    0.490
## Mining         -0.0641777  0.0478345  -1.342    0.187
## Manufacturing   0.0008496  0.0435787   0.019    0.985
## Government      0.0150582  0.1337220   0.113    0.911
## Recreation     -0.0533609  0.0608800  -0.876    0.386
## Nonspecialized  0.0338482  0.0419321   0.807    0.424
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.049 on 39 degrees of freedom
## Multiple R-squared:  0.1832, Adjusted R-squared:  0.05754
## F-statistic: 1.458 on 6 and 39 DF,  p-value: 0.2181
```

This shows a low R-squared value again.

For dependence on urbanization we have

```
##
## Call:
## lm(formula = Votes ~ Urbanization, data = urbanization1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.8947  -3.8916   0.0254   2.9724  16.8324
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.84803    1.20383   8.181 2.22e-10 ***
## Urbanization  0.03411    0.01602   2.130   0.0388 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6 on 44 degrees of freedom
## Multiple R-squared:  0.09345,    Adjusted R-squared:  0.07285
## F-statistic: 4.536 on 1 and 44 DF,  p-value: 0.03882
```

This is again very low forcing us to turn to regression analysis again.

For race dependence we obtain

```
##
## Call:
## lm(formula = Votes ~ White + Black + Native + Asian + Islander +
##     Mixed + Hispanic, data = race1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -11.8497  -2.2904  -0.0699   3.5542  15.8905
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  471.7418  1397.3312   0.338   0.7375
## White         -4.5590    13.9724  -0.326   0.7460
## Black         -4.5598    13.9485  -0.327   0.7455
## Native        -4.4472    13.7726  -0.323   0.7485
## Asian         -4.3554    13.9214  -0.313   0.7561
## Islander      -4.5901    13.7245  -0.334   0.7399
## Mixed         -5.4765    14.3690  -0.381   0.7052
## Hispanic      -0.2244     0.1156  -1.942   0.0595 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.992 on 38 degrees of freedom
## Multiple R-squared:  0.2192, Adjusted R-squared:  0.07537
## F-statistic: 1.524 on 7 and 38 DF,  p-value: 0.1887
```

This is again low justifying our use of regression trees.

Finally let us ook at the income variable.

```
##
## Call:
## lm(formula = absolute_votes ~ Income, data = full_data1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.7179  -3.3535   0.2397   3.8099  17.4382
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 10.40580    1.87780   5.541 1.59e-06 ***
## Income       0.05026    0.06957   0.722    0.474
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.264 on 44 degrees of freedom
## Multiple R-squared:  0.01172,    Adjusted R-squared:  -0.01074
## F-statistic: 0.5219 on 1 and 44 DF,  p-value: 0.4738
```

This has a low R-squared value.

Thus we use regression analysis to complete our study. This will be presented in the final report.

Limitations

This problem is merely exploratory and not predictive. We do not attempt to predict the outcome of the general elections based on the results of the preliminary. We could take a shot at this by making a comparison between the voters of Hillary Clinton and Donald Trump. We could then provide a strategy as to how one can win the other's voters by saying which characrteristics they should target. This implies that the candidate has to change their election campaign suitably to win the desired voters' endorsement.