BU425 - Assignment 1

Salar Ghamat

Oct. 29, 2021

By - Evan Friedlan

**Executive Summary**

The 2008 Democratic Primaries were a battle between Barack Obama and Hillary Clinton. Midway through, they were locked in close competition and presidential candidacy was still up for grabs. We will go back in time to the winter of 2008 and use US Census Data to analyze how the campaigns have been going, and who is expected to win. By building predictive models, we are able to project results the upcoming primaries and explain the major factors in Obama’s and Clinton’s success. This information would have been valuable to either candidate and this report will recommend strategies for both based on the results therein*. All RStudio code that was used has been included as an R file in the assignment submission, for your reference.*

On February 19, 2008, the U.S. Democratic primaries were in full swing. Barack Obama had just taken the lead when his primary opponent, Hillary Clinton, began running ads in Ohio aimed at middle-class, blue-collar voters. One of these ads “Night Shift”, seemed disingenuous to those voters. Obama had also had some missteps, like when he gave a speech about Whole Foods to farmers in Iowa. With 1,131 of 2,868 counties left to go, the candidates and their teams need to know how they are expected to fare in the upcoming states’ primaries and caucuses and where they should focus their efforts. With limited resources, they will want to allocate their capital and time to the most important voters that may swing the election. This report will predict their success in each county and overall chance of victory, as well as illuminate whether Clinton’s “Night Shift” ad and Obama’s “Down on the Farm” speech were well positioned and targeting the right voters.

Access to extensive demographic data from the U.S. Census Bureau was provided. This data contains categories such as average income, land area, ethnicity breakdowns, and much more. To prepare the data for analysis, I first replaced any NA (unavailable data) values in the dataframe with the mean value for each column. This assuredly skewed the data somewhat but is unlikely to make a huge difference in the predictive model as columns with many NAs were not included in the forecast for victory. The next adjustment was to change the type of Region and ElectionDate categories to factor and date, respectively. The data was then split into two sets, a training set for rows before February 19th, 2008, and a testing set for that day and after (See Technical Discussion T0 for more insight). Lastly, two subsets of the training set were created, election\_data\_smaller\_train with 75% of the train set sampled and election\_data\_validation with the other 25%. This was later used to check for overfitting and biased data. It would have also been useful to have a dataframe containing information on farmers’ support before the “Down on the Farm” speech. Comparing this with the actual results of primaries in counties with high FarmArea to LandArea ratio would help determine the effectiveness of Obama’s speech.

To analyze the data, I broke the problem into its two parts: forecasting election results and analyzing speech positioning. We will start with the election results. The first thing I did was create a correlation matrix (Technical Discussion T1). With this, I could see which variables affected primaries’ outcomes and which variables seemed to measure similar things (Appendix A1). After confirming the correlations (TD T2) I was left with a model, election\_linear\_model\_4, that was far too complex but a good jumping off point. Before making any more changes to the model, I checked its train error. This gave me a LogitTrainError of 20.8%. The same process (TD T3) was used on the 25% sample that the model had not yet seen, resulting in a LogitTestError of 20.7%. This is a very good result. The errors are low and close together, therefore it appears to be generalizable to unseen data. This model is still too complex, with 15 variables (See Appendix A2) meaning that it is likely to overfit on data that is not as similar as the two training sets. Before simplifying the model, however, I checked to see if the cut of 0 was reasonable to use, and it was (TD T4), (Appendix A3). Then, I took the most significant variables that correlated with Obama\_margin\_percent (O\_M\_P) and created a final linear model (Appendix A6). After testing for error, we find that the model predicts 22.8% for the training set and 19.7% for the validation set (Appendix A7). When looking at the model summary, we see that all variables are extremely powerful and the adjusted R-squared is 0.643 (TD 7 for how I came to use these variables). This is not much smaller than the complex model’s 0.697 with much more complexity. This suggests that our final model has few variables that are extremely powerful predictors. The residuals were generally homoscedastic (Appendix A8). I also tried to use a GLM and a regression tree to model the primaries’ results but found that the linear model was the most effective (Refer to the Technical Discussion for further reading). The GLM used Obama\_wins (O\_W) as my classifier (TD T5). It makes more sense to use the linear model over this GLM, because we want to also see by how much a candidate will win. The last type of model I tried out were regression trees with O\_W (TD T6), and a second version with O\_M\_P (Appendix A12). Then testing the train and test error for the second, it achieved 25.7 and 24.1 percent, (Appendix A13). This higher error led me to select election\_linear\_model\_final for forecasting. With the model decided, I could now run my linear on the real test data: the primaries that had not yet taken place. Testing first for accuracy, I found that the Mean Error was very close to zero for both training and validation sets (Appendix A14). This suggests that the model is extremely accurate, at least on the data before the 19th. Finally, I ran the model on the unseen, undecided counties. (Appendix A15).

The results suggest that Obama has his work cut out for him, with a current projection of 770 losses to 361 wins. With the reduced rate of Black and high school graduated voters in the upcoming primaries, it appears unlikely that an Obama victory will occur. These two factors are extremely significant in Obama’s success. It also appears that his appeal to farmers had little effect on voters. Even if some farmers were swayed by his awkward “Down on the Farm” speech, we can see in Appendix A16 that percent of farm area has very little, if any, correlation with Obama’s wins. This does not bode well for Mr. Obama, and he will need to do something drastic if he wants to stand a chance. I would recommend speeches and ads that appeal outside of his current demographic. Conversely, Ms. Clinton’s new ad seems to be targeting a potentially significant voter base. Blue collar workers seem to have a strong impact on results (TD 8), (Appendix A17), and blue collar workers tend to vote for Clinton more (TD 9). Also, in the undecided counties there is a greater frequency of blue-collar workers than the training set. Clinton should continue to hold strong to her current slight lead of 51% (Appendix A18) over Obama but be careful to not make any missteps that may turn off voters. If she can maintain and extend her lead, Clinton should have no problem securing the victory.

This report has demonstrated the power of predictive modelling in achieving real-world results. Whether being used to help Obama or Clinton, each candidate now knows their current position in the race and their expected results in the upcoming primaries. With this knowledge, they can adjust their strategies to overcome any previous missteps in speeches or rhetoric and stand a better chance at winning.

Appendix

A1 – Correlation Matrix

Chart, scatter chart

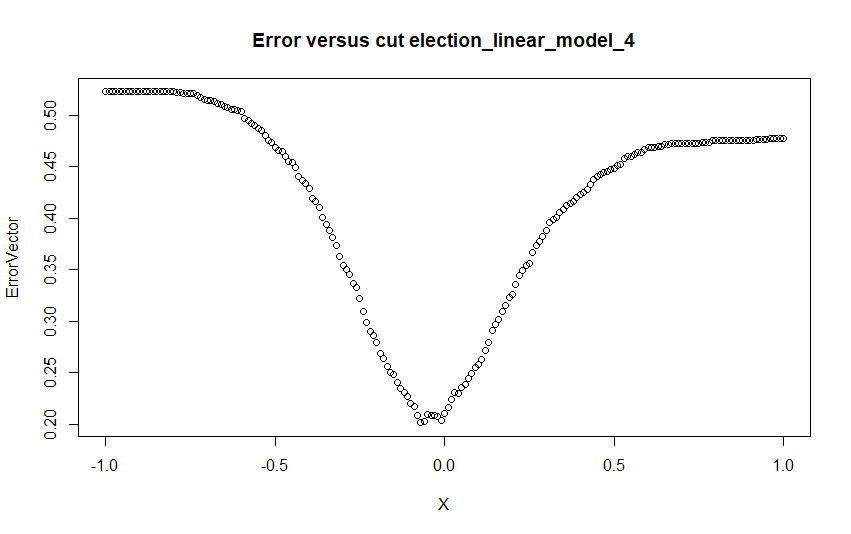
Description automatically generated

A2 – Complex Linear Model After Stepping

Table

Description automatically generated

A3 – Error vs Cut for Complex Model



A4 – Final Linear Model (Including PopDensity)

Text, table

Description automatically generated

A5 – Outliers caused by PopDensity Inclusion

Chart, line chart

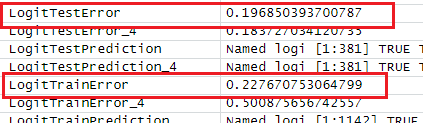
Description automatically generated

A6 – Final Linear Model (No PopDensity, to be used in Forecasting)

Text, table

Description automatically generated with medium confidence

A7 – Validation (Test) and Train Error of Final Linear Model



A8 – Residuals vs Fitted for Final Model

Chart, scatter chart

Description automatically generated

A9 – Error vs Cut for Final Model

Chart, scatter chart

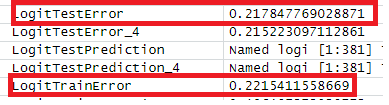
Description automatically generated

A10 – Generalized Linear Model

Table

Description automatically generated

A11 – GLM Validation & Train Error

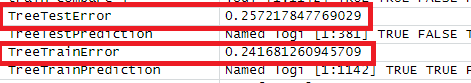


A12 – Regression Tree Obama\_margin\_percent (Without PopDensity)

Diagram

Description automatically generated

A13 – Tree Validation & Train Error (Without PopDensity)



A14 – Re-Test of Accuracy of Final Linear Model on Train & Validation Sets

Text, letter

Description automatically generated

Text

Description automatically generated

A15 – Final Projected Campaign Forecast

Text

Description automatically generated with medium confidence

A16 – Farm Area / Land Area Correlation with Obama\_margin\_percent

Scatter chart

Description automatically generated with low confidence

A17 – Blue Collar Worker Model

Text, letter

Description automatically generated

A18 – Clinton’s Lead as of February 19th



Technical Discussion

T0

* To the training set, which contains votes already cast, columns Other\_Candidates and Other\_Candidates\_Percent were added and used to cull rows that had 20% or more votes cast for candidates other than Obama or Clinton. Also, columns Obama\_margin, Obama\_margin\_percent, and Obama\_wins were created. This takes TotalVotes and Obama & Clinton votes to see where and by how much Obama won or lost each county. These columns will be used as the target in our model to predict results.

T1

* Instead of using the pairs() function, I created a correlation matrix using corrplot(). This included all variables from the data including Obama\_margin\_percent, Obama\_wins, etc.

T2

* To confirm the matrix’s insights, I ran a backward stepwise function and used variance inflation factor on a linear model containing every column of the election\_data\_smaller\_train frame. Removing insignificant and uncorrelated variables and repeating this process a couple of times,

T3

* Using a cut value of 0 (negative values of Obama\_margin\_percent = loss, positive = win), I compared the predicted values of model 4 with the actual values of the training data (75% sample).

T4

* Logically, it makes sense to place the border of win and loss at a margin of 0, but the model may be more accurate with another value; so, we have to check. Using a for loop to run through values of cut between -1 and 1, we can see that the error is in fact lowest at around 0, therefore we will leave it.

T5

* In this case, Obama\_wins will be used over Obama\_margin\_percent because the model requires a variable that has two potential answers. We use the election\_glm to predict the probability that a given county will be a win or loss (1 or 0) for Obama. Using similar testing methods as previous, I found that the same variables were still the most effective in prediction using the GLM (Appendix A10). Error testing showed the accuracy to be 21.78% and 22.15% on the train and validation sets, respectively (Appendix A11). I discovered similar results as election\_linear\_model\_final, which makes sense as it is calculating a similar prediction using the same variables

T6

* This type of model can split data into branches to classify predictions into two categories, so I used Obama\_wins. Once again, the same four variables were used, because I found that including PopDensity in this model caused a significant change to the difference in Train and Test Error. It caused the error rate to be larger on unseen data, supporting my earlier hypothesis that it did not contribute much to prediction and increased risk of overfitting.

T7

* I first tried to include PopDensity, but it was the one variable that had less significance (Appendix A4). I also tried switching in MedicareRate and SocialSecurityRate instead because they were correlated with PopDensity. What I found, however, was that the error rates did not have any noticeable changes, and the VIFs of the model rose slightly. The only improvement was found by removing PopDensity altogether. This kept the error rates nearly identical and less potential for overfitting. I thought that this might make the model underfit, but the PopDensity was introducing too much error with outliers like New York City (Labelled 1135 in Appendix A5).

TD 8

* Blue-collar workers are represented in the model by a lack of Bachelors degree and percent of employment in Manufacturing

TD 9

* Higher ManfEmployment and lower Bachelors are negatively correlated with Obama’s chances of winning, demonstrating that.