Segmenting Clinton & Obama Voters

Assignment 1

Paris Ngow - 180281890

BU425

Salar Ghamat

October 23rd, 2022

Table of Contents

Executive Summary

The Democratic party presidential primaries were currently taking place and the winner stood a good chance of becoming the next president of the United States. Thus, there was interest in predicting who would win between the party's two main candidates, Barak Obama, and Hilary Clinton.

An analysis was conducted using data from the 2007 U.S. Census Bureau, which contained demographic information about the counties. The data was cleaned and formatted to be used for this analysis. A column containing Obama's win margin percentage was created for a linear regression model to predict, while a binary column indicating if Obama would win was created for a classification decision tree model to predict. Clinton's results would be the negative margin percent and inverse binary value.

Of the counties left to vote, the results from the linear regression model indicated that Clinton would win whereas the decision tree model indicated that Obama would win. However, these models were not able to consider the fact that each candidate was also conducting a campaign to convince the remaining citizens to vote for them. So, the analysis also looked at whether each candidate's current campaign strategy would be effective in winning over a majority of the remaining voters.

Clinton's campaign, which was targeted at middle-class, blue-collar workers, might be more effective within a higher-income group as it showed more significance in predicting her win margin percentage. Furthermore, only a small number of citizens worked in the blue-collar sector of manufacturing and employment in the sector held little predictive significance. However, the data that was provided wasn't comprehensive enough to detail the employment in the other blue-collar sectors.

Obama elected to target farmers in the United States, which appeared to be a smart decision from a simple analysis of the data. However, when using that data to predict which candidates would win, it was clear that farmers and their farmland did not have a positive relationship with a successful election. Thus, it is recommended that Obama shift his focus to a more generalized campaign, rather than pile all his eggs into appealing to farmers who were not likely to help his case.

Problem Definition

The date is February 19th, 2008, and the delegate count for the Democratic Party's presidential primaries was quickly ending. The race for votes was extremely close between Barak Obama and Hilary Clinton. The Democratic party was interested in predicting which candidate would win the election and each candidate's margin of success. They also wanted to evaluate the candidates' campaign strategies and how they may affect their performance in the remainder of the election. This information was important to the Democratic party as the winner of the election could become the president of the United States.

Description of Data

The dataset that was received to conduct this analysis contains data collected from the U.S. Census Bureau's 2007 report. It consists of demographic information about the population in each county of the United States. It also includes how many citizens voted for Clinton and Obama. After an inspection of the dataset, the columns containing the state, region, and election type were all converted into factors. Several numeric columns were stored as characters, so they were also converted to numeric value types. Several columns also contained missing data points. All these columns were numeric, so we were able to reconcile the missing data by imputing them with the mean value of the respective columns.

To determine which candidate will win the presidential primaries, we created two target variables to predict – a continuous variable with the percentage margin of the candidate's win for regression and a binary variable indicating whether a candidate will win for classification. We created the variables for one candidate (Obama) as the other would hold the inverse results (Clinton's results would be a negative margin percent & 0 for the binary win variable). For simplicity, we will assume that there are only two candidates. We also split the data into training, testing, and validation sets based on the election date since we only wanted the training/validation sets to contain the counties that have voted. The validation set was obtained by randomly selecting 25% of the records from the training set. The testing set does not have labels with which to compare our models' predictions, so we cannot use it to evaluate them on unseen data. But we will use it to make a final prediction about the candidate who wins.

Analysis and Results

First, we will construct a linear regression (LR) model, which will predict the winning margin percentage, and a decision tree (DT) model to determine whether a candidate will win. To determine a win, the candidate must meet the threshold of obtaining over 50% of the votes in a county, a state, and the country. As it stands, Obama has won 17/30 states and Clinton will need to win 11/16 remaining states to make a comeback (See Appendix A). We excluded the columns in Appendix B from each model.

For the LR model, we ran the step function backwards and ended up with the model in Appendix C. On the validation set, the model boasted a root mean squared error of 0.1748 and a mean squared error of 0.1327. These values are comparable to the metrics on the training data, so the model was able to perform consistently on unseen data. Furthermore, the values are relatively close to 0 compared to the validation set's range of values for the column (-0.81, 0.72). On the test set, the LR predicted that Clinton would win at an average margin of 0.1928 and secure 13/16 of the remaining states (See Appendix D).

The DT model was built by setting the complexity parameter to 0.00001 and selecting the smallest tree iteration that was one standard deviation away from the minimum cross-validation error to avoid overfitting (See Appendix E). On the validation set, the model performed with a sensitivity of 0.7817, specificity of 0.7913, and area under the curve of 0.786, meaning it was able to identify both classes correctly about 80% of the time. On the testing set, the model predicted that each candidate would win 8/16 of the remaining states (See Appendix F). With this result, Obama would win overall.

Despite the predictions about how things will end, the election is ongoing, and the efforts of each candidate's campaigns could still influence the outcome. Thus, we will evaluate each campaign's effectiveness. Clinton attempted to appeal to middle-class voters. If the mid-point of the middle class is the median income of all counties ($37,281.5), we can see that Clinton was targeting voters in the most populous income group of the remaining voters. (See Appendix G). However, median income, which was only notable in the LR model, indicated significance when the coefficient was negative (positive for Clinton). Thus, appealing to higher-income voters could improve Clinton's odds of winning. Clinton also

attempted to target blue-collar voters, but the median percentage of the remaining voters in each county who work in the manufacturing sector is about 10% (See Appendix H). By focusing on blue-collar voters, Clinton may only be able to capture a small amount of the remaining voters. But note that this only considered the manufacturing sector, and the remaining voters could be in other blue-collar sectors.

Obama attempted to appeal to farmers in the country who weren't benefitting from increased prices. A sizable amount of the remaining voters are located in the South and Midwest (where Iowa is located) and they are the two regions with the most farmland (See Appendix I & J). We can also see that Obama has been able to appeal to voters from both regions, mostly in the South (See Appendix K). Both models indicated that voters in the South would play a significant factor in predicting who would win, however, the LR model indicated that voters from the South would play less of a factor than voters in the Midwest. It also indicated that the amount of farmland would play an inverse role in predicting a winner (less is more significant). So, perhaps Obama should shift the focus of his campaign away from farmers.

Recommendation and Action Plan

The fact that both models predicted different candidates to win reminds us that the election is not set in stone. Thus, the effectiveness of the campaigns will be crucial to influence votes. For Obama to increase his chances of winning, he should focus on Midwest states and shift to emphasize that groceries are becoming unaffordable instead of appealing to farmers. The LR model indicated lower incomes were more significant, and it would appeal more to the general population. Clinton, on the other hand, should consider targeting higher-income voters and obtain more data to see whether it is worthwhile to continue appealing to blue-collar workers, as employment in the manufacturing sector held little significance.

Conclusion

As it stands, the LR model was able to predict that Clinton would win the Democratic presidential primaries with an average margin percentage of 0.1928 in the remaining states, while the DT model indicated Obama would win. However, historical data is not the only factor in determining how the remaining citizens will cast their votes since each candidate’s campaign could still sway their decisions.

Appendices

Appendix A – Current Voter Standings

States	Percentage of Votes	Win (Obama = 1, Clinton = 0)
AL	0.52238806	1
AR	0.04000000	0
AZ	0.13333333	0
CA	0.34482759	0
CO	0.78125000	1
CT	0.75000000	1
DE	0.66666667	1
FL	0.13432836	0
GA	0.69811321	1
IA	0.51515152	1
ID	0.86363636	1
IL	0.86274510	1
LA	0.68750000	1
MA	0.35714286	0
MD	0.66666667	1
ME	0.75000000	1
MN	0.79310345	1
MO	0.05217391	0
NE	0.32258065	0
NH	0.50000000	0
NJ	0.23809524	0
NM	0.18181818	0
NV	0.64705882	1
NY	0.01612903	0
OK	0.01298701	0
SC	0.95652174	1
TN	0.08421053	0
UT	0.79310345	1
VA	0.74626866	1
WA	0.58974359	1

Appendix B – Variables Excluded from The Models

Variable	Model Left Out Of	Reason Left Out
Obama_margin	Both	Would not be available when predicting the margin for counties who haven't voted
Obama_margin_percent	Decision Tree	Would not be available when predicting the margin for counties who haven't voted
Obama_wins	Linear Regression	Would not be available when predicting the margin for counties who haven't voted
Obama	Both	Would not be available when predicting the margin for counties who haven't voted
Clinton	Both	Would not be available when predicting the margin for counties who haven't voted

TotalVote	Both	Would not be available when predicting the margin for counties who haven't voted
ElectionDate	Both	Dates from training set would not be found in the testing data (as we split based on it)
County	Both	Will have a unique value for each record, useless for model to find a pattern and use to estimate & would not be available when predicting the margin for counties who haven't voted
State	Both	Would not be available when predicting the margin for counties who haven't voted
FIPS	Both	Will have a unique value for each record, useless for model to find a pattern and use to estimate

Appendix C – Linear Regression Model

```
Call:
lm(formula = Obama_margin_percent ~ Region + ElectionType + Age35to65 +
    Age65andAbove + White + Asian + AmericanIndian + Hawaiian +
    Bachelors + Poverty + MedianIncome + AverageIncome + UnemployRate +
    ManfEmploy + MedicareRate + SocialSecurity + RetiredWorkers +
    Disabilities + DisabilitiesRate + Homeowner + SameHouse1995and2000 +
    PopDensity + LandArea + FarmArea, data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-0.85675 -0.09448  0.00042  0.10725  0.52248

Coefficients:
                          Estimate   Std. Error t value             Pr(>|t|)
(Intercept)           2.225196921  0.128370672  17.334 < 0.0000000000000002 ***
RegionNortheast      -0.199337336  0.021762797  -9.160 < 0.0000000000000002 ***
RegionSouth          -0.243649523  0.015476184 -15.744 < 0.0000000000000002 ***
RegionWest            0.132843393  0.018372859   7.230  0.00000000000082762 ***
ElectionTypePrimary  -0.113019252  0.014356071  -7.873  0.00000000000000738 ***
Age35to65            -0.003768053  0.001777369  -2.120             0.034197 *
Age65andAbove        -0.013962206  0.002589516  -5.392  0.00000008303270583 ***
White                -0.019855528  0.000490845 -40.452 < 0.0000000000000002 ***
Asian                -0.031539003  0.003059479 -10.309 < 0.0000000000000002 ***
AmericanIndian       -0.018157194  0.001021235 -17.780 < 0.0000000000000002 ***
Hawaiian             -0.086863310  0.043415995  -2.001             0.045634 *
Bachelors             0.014671448  0.001122739  13.068 < 0.0000000000000002 ***
Poverty              -0.021648829  0.002885113  -7.504  0.00000000000011601 ***
MedianIncome         -0.000004572  0.000001339  -3.414             0.000661 ***
AverageIncome        -0.000002635  0.000001464  -1.800             0.072055 .
UnemployRate         -0.016130919  0.003828600  -4.213  0.00002694129592835 ***
ManfEmploy           -0.001607240  0.000725764  -2.215             0.026967 *
MedicareRate          0.000006486  0.000001679   3.863             0.000118 ***
SocialSecurity       -0.000007383  0.000002631  -2.807             0.005083 **
RetiredWorkers        0.000008886  0.000003747   2.371             0.017871 *
Disabilities          0.000005503  0.000001709   3.219             0.001317 **
DisabilitiesRate     -0.000034156  0.000006482  -5.269  0.00000016068071738 ***
Homeowner             0.002064420  0.001095891   1.884             0.059822 .
SameHouse1995and2000  0.005211502  0.001056104   4.935  0.00000090907329131 ***
PopDensity           -0.000007282  0.000003028  -2.405             0.016331 *
LandArea              0.000006355  0.000003758   1.691             0.091064 .
FarmArea             -0.000059131  0.000016733  -3.534             0.000424 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1628 on 1275 degrees of freedom
Multiple R-squared:  0.7103,    Adjusted R-squared:  0.7044
F-statistic: 120.2 on 26 and 1275 DF,  p-value: < 0.00000000000000022


 [1] "Min, Max of Obama_margin_percent"
 [1] -0.8093184
 [1] 0.7260982
 [1] "Training Set"
                                      ME      RMSE       MAE MPE MAPE
 Test set 0.000000000000000000003989252 0.1611283 0.1246789 NaN  Inf
 [1] "Validation Set"
                       ME      RMSE       MAE  MPE MAPE
 Test set -0.004178193 0.1747875 0.1326621 -Inf  Inf
```

Appendix D – Linear Regression Results on Testing Data

States	Percentage of Votes	Win (Obama = 1, Clinton = 0)
HI	0.00000000	0
IN	0.16304348	0
KY	0.00000000	0
MS	0.70731707	1
MT	0.73214286	1
NC	0.45000000	0
OH	0.27272727	0
OR	0.30555556	0
PA	0.08955224	0
RI	0.20000000	0
SD	0.22727273	0
TX	0.01593625	0
VT	0.21428571	0
WI	0.22222222	0
WV	0.00000000	0
WY	1.00000000	1

Appendix E – Decision Tree Model

	CP	nsplit	rel error	xerror	xstd
1	0.2588997	0	1.00000	1.00000	0.029156
2	0.1197411	1	0.74110	0.75243	0.027977
3	0.0857605	2	0.62136	0.71845	0.027678
4	0.0412621	3	0.53560	0.61974	0.026605
5	0.0194175	5	0.45307	0.52427	0.025243
6	0.0113269	6	0.43366	0.49353	0.024729
7	0.0080906	10	0.38350	0.48706	0.024615
8	0.0072816	13	0.35599	0.48706	0.024615
9	0.0064725	15	0.34142	0.48220	0.024529 *
10	0.0053937	16	0.33495	0.48706	0.024615
11	0.0048544	19	0.31877	0.48220	0.024529
12	0.0043150	21	0.30906	0.48544	0.024587
13	0.0032362	24	0.29612	0.51456	0.025085
14	0.0028317	32	0.27023	0.50971	0.025005
15	0.0021575	37	0.25566	0.52589	0.025269
16	0.0016181	40	0.24919	0.52913	0.025321
17	0.0000100	41	0.24757	0.54854	0.025622

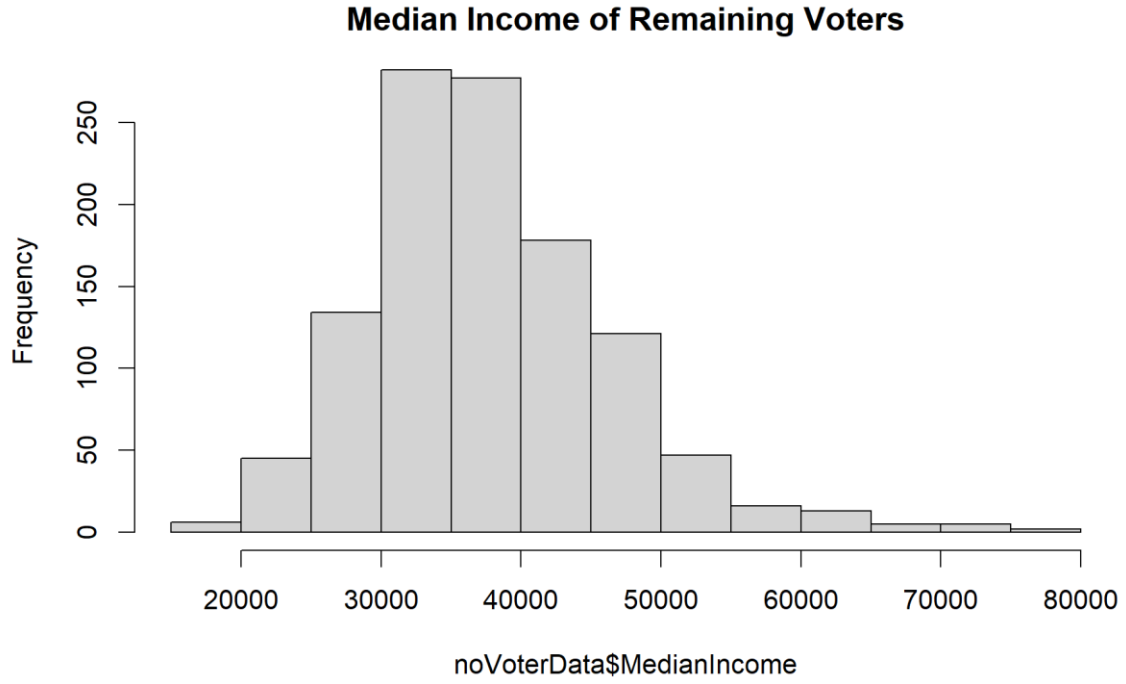
* → Minimum cross validation error
The circled record is the smallest tree iteration within 1 xstd from the minimum cross validation error

```
Node number 89: 28 observations
  predicted class=1  expected loss=0.2857143  P(node) =0.02150538
    class counts:     8    20
   probabilities: 0.286 0.714


Classification tree:
rpart(formula = Obama_wins ~ . - Obama_margin - Obama_margin_percent -
    Obama - Clinton - TotalVote - ElectionDate, data = train,
    method = "class", cp = 0.0113269)

Variables actually used in tree construction:
[1] Age35to65          Black              HighSchool         Region
SocialSecurityRate

Root node error: 618/1302 = 0.47465

n= 1302

        CP nsplit rel error  xerror     xstd
1 0.258900      0   1.00000 1.00000 0.029156
2 0.119741      1   0.74110 0.75081 0.027963
3 0.085761      2   0.62136 0.65858 0.027065
4 0.041262      3   0.53560 0.60841 0.026461
5 0.019417      5   0.45307 0.50647 0.024950
6 0.011327      6   0.43366 0.49353 0.024729
Warning: may not be applicable for this method

                  Reference
          Prediction   0   1
                   0 179  43
                   1  50 163

                  Accuracy : 0.7862
                    95% CI : (0.7446, 0.8238)
       No Information Rate : 0.5264
       P-Value [Acc > NIR] : <0.0000000000000002

                     Kappa : 0.5719

    Mcnemar's Test P-Value : 0.5338

               Sensitivity : 0.7817
               Specificity : 0.7913
            Pos Pred Value : 0.8063
            Neg Pred Value : 0.7653
                Prevalence : 0.5264
            Detection Rate : 0.4115
      Detection Prevalence : 0.5103
         Balanced Accuracy : 0.7865

          'Positive' Class : 0

Setting levels: control = 0, case = 1
Setting direction: controls < cases
Area under the curve: 0.7865
```

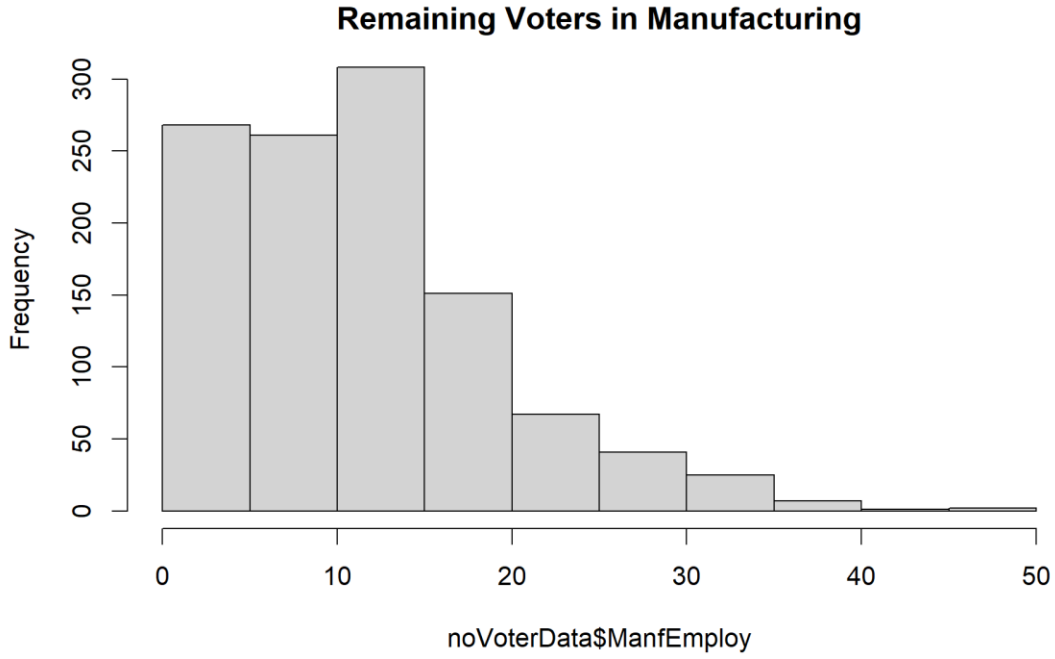
Appendix F – Decision Tree Results on Testing Data

States	Percentage of Votes	Win (Obama = 1, Clinton = 0)
HI	0.75000000	1
IN	0.78260870	1
KY	0.03333333	0
MS	0.81707317	1
MT	0.55357143	1
NC	0.50000000	0
OH	0.81818182	1
OR	0.52777778	1
PA	0.08955224	0
RI	0.40000000	0
SD	0.42424242	0
TX	0.05976096	0
VT	0.42857143	0
WI	0.73611111	1
WV	0.07272727	0
WY	0.78260870	1

Appendix G – Histogram Outlining the Median Income of the Remaining Voters

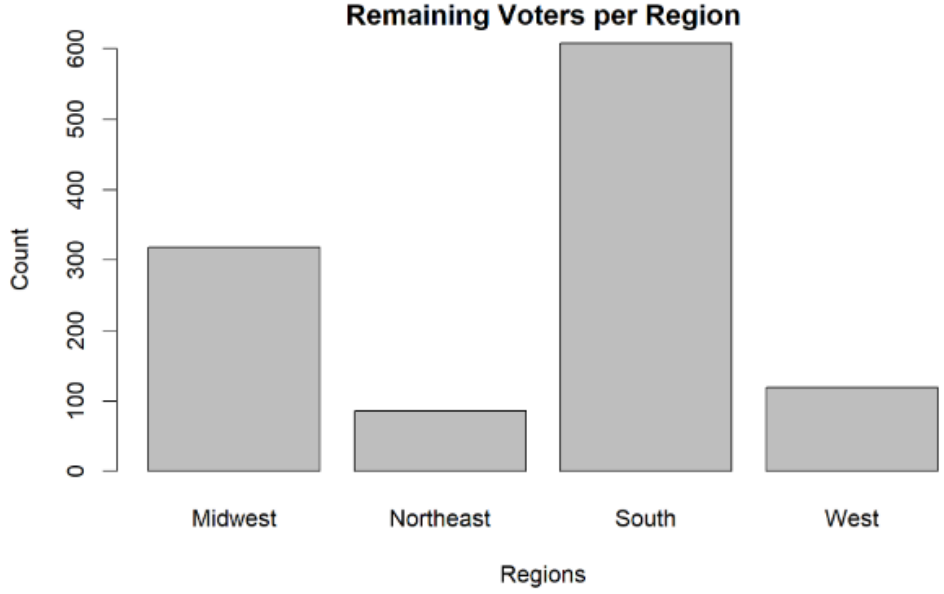


Appendix H - Histogram Outlining the Percentage of the Remaining Voters in Manufacturing Jobs



	[,1]
Min	0.2024292
First Quartile	5.1739179
Median	10.6564166
Third Quartile	15.3142716
Maximum	30.4573459

Appendix I – Bar Chart Containing the Number of Remaining County Votes per Region



Appendix J – Farmland Owned by Remaining Voters per Region

Region	Sum of Farmland (square miles)	Sum of Land Area (square miles)	Percent Farmland
Midwest	89,465.652	223,849	39.97%
Northeast	9,634.303	57,218	16.84%
South	165,818.258	432,718	38.32%
West	112,540.652	352,906	31.89%

Appendix K – Bar Chart Containing the Number of Obama's County Votes per Region