DUKE
THE FUQUA SCHOOL OF BUSINESS

2008 Democratic Primaries – Clinton vs. Obama[i]

Background

It was February 19, 2008. One week earlier, Barack Obama had taken the lead in the delegate count during the Democratic Party's presidential primaries, the winner of which would face the Republican Party's nominee in the general election to become the next president of the United States (POTUS).

On that day in February, Hillary Clinton, Obama's primary opponent, began running ads in Ohio aimed at middle-class, blue-collar voters. One ad, "Night Shift," closed showing Clinton at her desk: "She understands. She's worked the night shift, too." But had Clinton ever worked the night shift? Her spokesperson said it was a "rhetorical reference" to working late nights as a lawyer, First Lady, and senator [1].

Clinton was not alone in her awkward appeals to voters in key demographics. Months earlier at a campaign stop in Iowa, Obama noted that while produce prices had risen in grocery stores, farmers had not benefited from increases in crop prices: "Anybody gone into Whole Foods lately and see what they charge for arugula? I mean, they're charging a lot of money for this stuff." At the time, there wasn't a single Whole Foods in the state of Iowa [2].

What did these missteps say about the candidates' campaign strategies? Were they targeting the right voters? Had they crafted the right messages for these segments? One key in answering these questions was to analyze votes that had already been cast.

Demographic Data

In November 2007, the U.S. Census Bureau issued its County and City Data Book [3] which contained extensive demographic information by state, county, and city. The Census Bureau grouped the 50 U.S. states into four regions (see Exhibit 1). Within the United States, there were a total of 3,141 counties.

On its website, the Census Bureau released the 2007 data tables from the County and City Data Book. These tables contained information commonly used by marketers to segment a population—gender, age, race, ethnicity, education, income, employment status, knowledge of languages, government dependency, disabilities, home ownership, mobility, population, population density, and region [4]. Key demographic data by county were extracted from these tables and placed in a spreadsheet file.

Vote Data

Candidates for the Democratic nomination won delegates—both pledged delegates and superdelegates—through a complicated process followed closely by many news organizations and depicted in many different ways.[1] On its website, CNN displayed an interactive graphic where candidates were depicted as donkeys in a foot race; visitors could roll a cursor over a candidate's donkeys and the current delegate total would appear [5].

Politicians cared about votes the way marketers cared about sales. Because voting results dictated delegate commitment, major news outlets carefully tracked totals by county. A visitor had access to county vote data (also included in the spreadsheet) by rolling a cursor over a state's map.[2]

As of February 19, 2008, of the 2,868 total counties for which county-level vote data would eventually become available, there were 1,131 counties left to report. An estimated 2,490 delegates were already committed, and 2,118 were needed to secure the party's nomination; there were still 1,744 delegates to be awarded in upcoming states' primaries and caucuses. (See Exhibit 2 for a list of past and upcoming primaries and caucuses.) Later that evening, the results of Hawaii's caucus and Wisconsin's primary would be announced. In these two states, and in over 1,000 counties in other states, it remained to be seen who would vote for Clinton and who would vote for Obama. Whoever won the most votes in these remaining primaries was likely to become the next POTUS.

What Followed and the Role of Analytics in Politics

Indeed, Obama won the Democrat nomination and won two Presidential elections. In the 2008 presidential election, Obama's targeters had assigned every voter in the country a pair of scores based on the probability that the individual would perform two distinct actions that mattered to the campaign: casting a ballot and supporting Obama.

For each battleground state every week, the campaign's call centers conducted 5,000 to 10,000 so-called short-form interviews that quickly gauged a voter's preferences, and 1,000 interviews in a long-form version that was more like a traditional poll. To derive individual-level predictions, algorithms trawled for patterns between these opinions and the data points the campaign had assembled for every voter—as many as one thousand variables each, drawn from voter registration records, consumer data warehouses, and past campaign contacts.

Furthermore, the Obama 2008 campaign also used analytics to increase donations using A/B testing.

---

[1] Pledged delegates were committed to a candidate in proportion to the votes cast for that candidate in the state's primary or caucus. Superdelegates (almost 20% of the total number of delegates) were free to commit to any candidate at any time.

[2] Several primaries and caucuses were excluded from the county-level vote data set because either Obama's name did not appear on the ballot, vote data was not reported by county, or a U.S. county was not represented.

The campaign tried four buttons and six different media (three images and three videos). A full-factorial multivariate test was used to statistically test all the combinations of buttons and media against each other at the same time. Since there were four buttons and six different media that meant a total of 24 (4 x 6) total combinations to test. Every visitor to the splash page was randomly shown one of these combinations and we tracked whether they signed up or not.

The winning variation had a sign-up rate of 11.6%. The original page had a sign-up rate of 8.26%. That's an improvement of 40.6% in sign-up rate. What does an improvement of 40.6% translate into? Instead of having nearly 10 million people signed up for the campaign, the original page would have obtained only close to 7,120,000 signups. That is a difference of 2,880,000 email addresses. These additional email addresses are essentially responsible for 288,000 more volunteers and an addition $60millions in donations.

This created a new culture, and the Obama 2012 campaign also used data analytics and the experimental method to assemble a winning coalition vote by vote. In doing so, it overturned the long dominance of TV advertising in U.S. politics and created something new in the world: a national campaign run like a local ward election, where the interests of individual voters were known and addressed [6].

References

[1] Ariel Alexovich, "New Clinton Ad: 'Night Shift,'" *New York Times*, February 19, 2008.

[2] Jeff Zeleny, "Obama's Down on the Farm," *New York Times*, July 27, 2007.

[3] County and City Data Book: 2007, 14th ed. (Washington, DC: U.S. Census Bureau, 2007).

[4] *County and City Data Book: 2007*, "State and County Data Tables," available at http://www.census.gov/statab/ccdb/ccdbstcounty.html (accessed May 27, 2013.)

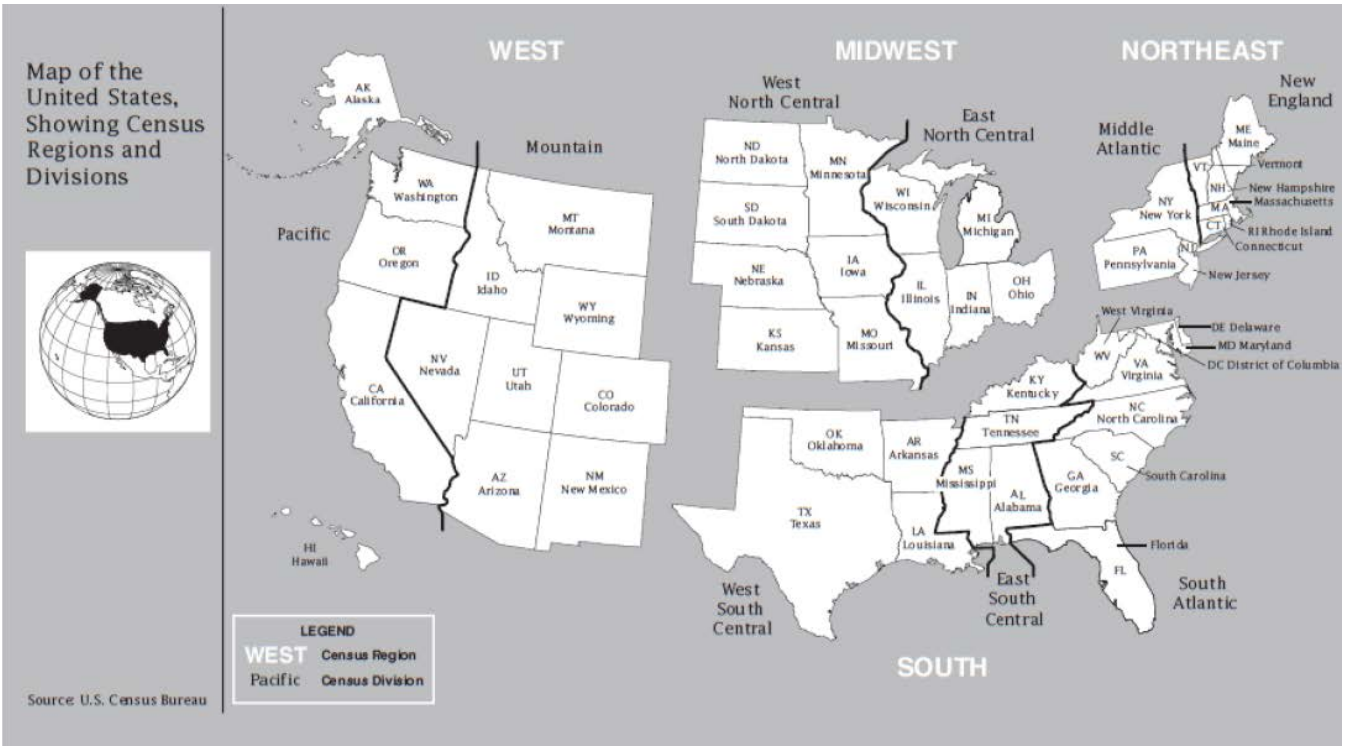
[5] "Election Center 2008: Results," CNNPolitics.com, http://www.cnn.com/ELECTION/2008/primaries/results /scorecard/#D (accessed May 27, 2013).

[6] How President Obama's campaign used big data to rally individual voters, MIT Technology Review, Dec 19 2012, http://www.technologyreview.com/featuredstory/509026/how-obamas-team-used-big-data-to-rally-voters/

[7] How Obama Raised $60 Million by Running a Simple Experiment, Optimizely Blog, November 29, 2010. https://blog.optimizely.com/2010/11/29/how-obama-raised-60-million-by-running-a-simple-experiment/

Exhibit 1 – Census Regions and Divisions

U.S. Census Bureau's Regions and Divisions for 2007



Source: *County and City Data Book: 2007*, 14th ed. (Washington, DC: U.S. Census Bureau, 2007).

Exhibit 2 – Past and Upcoming Democratic Party Primaries and Caucuses

Past Primaries and Caucuses				
Date	State Name	State Code	Election Type	Region
3-Jan	Iowa	IA	Caucus[a]	Midwest
8-Jan	New Hampshire	NH	Primary[b]	Northeast
15-Jan	Michigan*	MI	Primary	Midwest
19-Jan	Nevada	NV	Caucus	West
26-Jan	South Carolina	SC	Primary	South
29-Jan	Florida	FL	Primary	South
5-Feb	Alaska**	AK	Caucus	West
5-Feb	Alabama	AL	Primary	South
5-Feb	Arkansas	AR	Primary	South
5-Feb	American Samoa***	AS	Caucus	Territory
5-Feb	Arizona	AZ	Primary	West
5-Feb	California	CA	Primary	West
5-Feb	Colorado	CO	Caucus	West
5-Feb	Connecticut	CT	Primary	Northeast
5-Feb	Democrats aboard***	DA	Primary	Overseas
5-Feb	Delaware	DE	Primary	South
5-Feb	Georgia	GA	Primary	South
5-Feb	Idaho	ID	Caucus	West
5-Feb	Illinois	IL	Primary	Midwest
5-Feb	Kansas**	KS	Caucus	Midwest
5-Feb	Massachusetts	MA	Primary	Northeast
5-Feb	Minnesota	MN	Caucus	Midwest
5-Feb	Missouri	MO	Primary	Midwest
5-Feb	North Dakota**	ND	Caucus	Midwest
5-Feb	New Jersey	NJ	Primary	Northeast
5-Feb	New Mexico	NM	Primary	West
5-Feb	New York	NY	Primary	Northeast
5-Feb	Oklahoma	OK	Primary	South
5-Feb	Tennessee	TN	Primary	South
5-Feb	Utah	UT	Primary	West
9-Feb	Louisiana	LA	Primary	South
9-Feb	Nebraska	NE	Caucus	Midwest
9-Feb	Washington	WA	Caucus	West
10-Feb	Maine	ME	Caucus	Northeast
12-Feb	District of Columbia***	DC	Primary	South
12-Feb	Maryland	MD	Primary	South
12-Feb	Virginia	VA	Primary	South

Upcoming Primaries and Caucuses				
Date	State Name	State Code	Election Type	Region
19-Feb	Hawaii	HI	Caucus	West
19-Feb	Wisconsin	WI	Primary	Midwest
4-Mar	Ohio	OH	Primary	Midwest
4-Mar	Rhode Island	RI	Primary	Northeast
4-Mar	Texas	TX	Primary	South
4-Mar	Vermont	VT	Primary	Northeast
4-Mar	Texas	TX	Caucus	South
8-Mar	Wyoming	WY	Caucus	West
11-Mar	Mississippi	MS	Primary	South
9-Apr	Virgin Islands***	VI	Caucus	Territory
22-Apr	Pennsylvania	PA	Primary	Northeast
3-May	Guam***	GU	Caucus	Territory
6-May	Indiana	IN	Primary	Midwest
6-May	North Carolina	NC	Primary	South
13-May	West Virginia	WV	Primary	South
20-May	Kentucky	KY	Primary	South
20-May	Oregon	OR	Primary	West
1-Jun	Puerto Rico***	PR	Primary	Territory
3-Jun	Montana	MT	Primary	West
3-Jun	South Dakota	SD	Primary	Midwest

* Obama's name did not appear on the ballot. He boycotted the primary because it violated Democratic Party rules.

** Vote data not reported by county.

*** Demographic data not included in data set because the district/territory did not contain a U.S. county.

[a] Voting in a caucus is done openly by raising hands or breaking into groups.

[b] Primary voters cast secret ballots for the candidates of their choosing.

Data source: "Election Center 2008: Results," CNNPolitics.com, http://www.cnn.com/ELECTION/2008/primaries/results/scorecard/#D (accessed May 27, 2013).

Exhibit 3 – Variable definitions

Data Type	Column Headers in the Data sheet	Description
County	County	County
County	State	State
County	Region	Region
County	FIPS	Fedral Information Processing Standard (Unique ID for counties in the US)
County	ElectionDate	Date of Election
County	ElectionType	Type of Election (Primary or Caucus)
Vote	TotalVote	Total votes cast
Vote	Clinton	Votes cast for Clinton
Vote	Obama	Votes cast for Obama
Demographic	MalesPer100Females	Males per 100 females
Demographic	AgeBelow35	Population aged less than 35 years
Demographic	Age35to65	Population aged between 35 and 65 years
Demographic	Age65andAbove	Population aged more than 65 year
Demographic	White	Population White (percent)
Demographic	Black	Population Black or African American (percent)
Demographic	Asian	Population Asian (percent)
Demographic	AmericanIndian	Population American Indian and Alaska Native (percent)
Demographic	Hawaiian	Population Hawaiian and other Pacific Islander (percent)
Demographic	Hispanic	Population Hispanic or Latino origin (percent)
Demographic	HighSchool	High school graduate or higher (percent)
Demographic	Bachelors	Bachelor's degree or higher (percent)
Demographic	Poverty	Persons in poverty in 2004
Demographic	IncomeAbove75K	Households with income of $75,000 or more in 1999 (percent)
Demographic	MedianIncome	Median household income in 2005
Demographic	AverageIncome	Per capita income in 2005
Demographic	UnemployRate	Unemployment rate in 2006
Demographic	ManfEmploy	Percent of total employment in manufacturing in 2005
Demographic	SpeakingNonEnglish	Speaking language other than English at home in 2000 (percent)
Demographic	Medicare	Medicare program enrollment in 2005 - Number
Demographic	MedicareRate	Medicare program enrollment in 2005 - Rate per 100,000 persons
Demographic	SocialSecurity	Social Security program beneficiaries in December 2005 - Number
Demographic	SocialSecurityRate	Social Security program beneficiaries in December 2005 - Rate per 100,000 persons
Demographic	RetiredWorkers	Social Security program beneficiaries in December 2005 - Retired workers, number
Demographic	Disabilities	Supplemental Security Income program recipients in 2005 - Number
Demographic	DisabilitiesRate	Supplemental Security Income program recipients in 2005 - Rate per 100,000 person
Demographic	Homeowner	Owner-occupied housing units in 2000 (percent)
Demographic	SameHouse1995and2000	Residing in same house in 1995 and 2000 (percent)
Demographic	Pop	Population 2006 (July 1)
Demographic	PopDensity	Population per square mile of land area (2006)
County	LandArea	Total county area (square miles)
Demographic	FarmArea	Total land in farms in the county in 2002 (square miles)

Case: 2008 Democratic Primaries – Clinton vs. Obama

Instructions

This is a team assignment. Each member of the team receives the same grade. Submission is online. In order to be graded, you need to upload ***one PDF file*** *(no longer than 5 pages with font size 12pt) and your R script (this should be well commented and run without errors). Any additional material you judge relevant that complements your submission can be submitted as additional files.* ***Make sure that the section number, team number and all names of the team members are clearly listed in the top of the first page of the PDF file.*** *Late submissions (but submitted before in-class discussions) or inappropriately formatted cases will have points deducted. Missed cases are worth 0 points.*

Assignment

There are five questions. Respond to all of them. Your answers should be clear and provide unambiguous recommendations when asked. Please provide explanations for your answers and any outputs that you feel are needed to support your argument. Keep in mind that this is an open ended exercise. There is no exact model of reality as discussed in class. You will be evaluated on your modeling of the problem, judgment of which core task to use, appropriateness of the choice of algorithm, and taking all of that to data. (Therefore do not stress about trying to fine tune details.)

An R script is provided with some data cleaning and specific models for Question 4.

Questions

1. Pick two (or more) variables and attempt to show a relation between them via visualization. As discussed before, this requires one to formulate a question, and to communicate clearly a conclusion based on data visualization (specify the why, what, how). (Note that in this question it is not required that the relationship displayed relates to the election.)

2. Provide a model to predict the winning spread of Obama over Clinton measured as percentage of the total vote. Describe clearly the core task, briefly discuss all the models you compared, state which metric is being used to evaluate performance, and how did you chose a final model. Apply and report a K-fold cross validation to evaluate the performance of your chosen model. Based on your final model, predict the winning spread percentages for the test sample (provide the R code that generate your predictions).

	Vote	**Demographic/county data**
Training set Primaries and caucuses before February 19, 2008 (1,737 rows)	Available	Available
Testing set Primaries and caucuses on or after February 19, 2008 (1,131 rows)	Not Available (submit predictions)	Available

3. In order to explore the data, apply one unsupervised learning tool (e.g., k-means, principal component analysis), interpret and communicate briefly the output (e.g., clusters, latent features), and attempt to obtain insights.

4. Several sources have been reporting that the demographic composition of the US is changing which can definitely impact how campaigns will be run. In many states, the Hispanic population is growing at a faster pace than others. Looking ahead, provide an estimate for what would have been the average impact on the winning spread for Obama over Clinton (measured in percentage of total voters) had the Hispanic demographic been 5% larger? What if the Black demographic was 5% larger? Be careful to isolate the impact of the specific demographic change alone. (This question would be too open ended. In the R starter script we provide a "simple" model with 1771 variables to be used for which we will assume that the Conditional Independence Assumption (CIA) holds.)

Table 1
U.S. Population, Actual and Projected: 2005 and 2050

	2005	2050
Population (in millions)	296	*438*
Share of total		
Foreign born	12%	*19%*
Racial/Ethnic Groups		
White	67%	*47%*
Hispanic	14%	*29%*
Black	13%	*13%*
Asian	5%	*9%*
Age Groups		
Children (17 and younger)	25%	*23%*
Working age (18–64)	63%	*58%*
Elderly (65 and older)	12%	*19%*

Note: All races modified and not Hispanic; American Indian/ Alaska Native not shown. See "Methodology."
Source: Pew Research Center, 2008

5. Choose one candidate. What kind of advice (based on data analytics) would you provide to your candidate? For example, which voter segment to target with their campaign messages and why? Or, how to allocate resources (budget and volunteer time) across regions and why? How would you communicate such insights?

[i] This case was created based on the case UVA-QA-0807 which was prepared by Associate Professor Kenneth C. Lichtendahl Jr. and Rohit Gupta (MBA '13) at Darden School of Business.