

Date: March 29, 2004

To: League of Conservation Voters Education Fund

From: Anna Greenberg, Greenberg Quinlan Rosner Research
Jeff Hayes, Stratalys Research

- SmarTargeting Validation Study -

The validation study for the League of Conservation Voters Education Fund (LCVEF) was encouraging, but the results were mixed. All microtargeting efforts are still a work in progress and it makes sense that we will adjust our strategies in response to continued testing of the models. We continue to believe that the most important validation is "on the ground" validation tested against traditional targeting methods. In other words, we will not know with any certainty that this strategy is better than what we have done in the past unless we determine that the lists we pull are an improvement over the lists we would have pulled with other information (e.g., polling). Models continue to be predictions about behavior, but we need to gauge actual behavior (which is a high bar) in order to determine if these models are an improvement over traditional methods.

Primary Model

The environmental activist model did not perform up to expectations. We categorized 16 percent of the Southwest electorate as "environmental activists" in the data.^Σ The validation study sought to confirm that, based on SmarTargeting modeling, the most targetable 5% of the electorate would contain as much as 38 percent environmental activists. We in fact found that 25 percent were environmental activists. That is, we had projected an efficiency ratio of 2.38 times the general population incidence (i.e., $38/16 = 2.38$), but validation results suggest an efficiency closer to 1.56. This means that if we contact 100 people we would get 25 activists, instead of 38 activists as the model would have predicted. This means the model is still identifying potential activists, but not to the degree that we had hoped. While this is somewhat discouraging, the model certainly helps us find more environmental activists than if a list had been draw at random.

^Σ Note that this is an arbitrary categorization. That is, a 16% activist incidence may appear too high or too low to an observer, but this is not consequential for SmarTargeting modeling.

Environmental Activist Model	ST Top 5% Projection (N=255)	ST Top 5% Validation Result (N=200)	General Population (N=5000)	Traditional Voter File Targeting of Registered Voters (N=2237)
Anti-Enviros (0-29)	9%	12%	22%	17%
Passive Enviros (30-69)	54%	64%	62%	64%
Active Enviros (70-100)	38%	25%	16%	19%

Still, there is much that can and will be done to attempt to improve our current performance. We believe that the survey data we have gathered represents a solid foundation; we are confident in the original sampling and research design. We believe that changing the modeling process, as opposed to say needing a new survey, will improve the accuracy because the model performed well in helping us identify people attitudinally. In other words, voters largely had the attitudes we predicted them to have about the environment and other issues such as choice, corporate America, and the like. This suggests that we have the building block for a better model, but we need to improve the techniques used to create the model. The best way to do this is by gaining additional feedback from refining the modeling and through more sophisticated validation tests.

Attitude Questions	Projected	Validation
Endangered species	78.83	78.04
Environmental groups	69.12	66.02
Labor unions	64.47	60.67
Pro-life groups	48.88	46.80
Big corporations	42.58	45.89

There are a number of new modeling strategy alternatives that should be evaluated through additional validation. Data mining algorithms are known for "overfitting", which means they can create very accurate models to describe the data set being analyzed but these models may be based on patterns peculiar to that dataset and not found in the actual population of interest. Consequently, we will estimate alternate SmarTargeting models that will constrain overfitting even more than we attempted in our initial round of modeling. It is also possible that different algorithms may generate different levels of efficiency. These differences may be difficult to accurately evaluate when analyzing the dataset that the algorithms are themselves modeling; external validation is needed to adjudicate between them. Our initial modeling effort was a two-stage approach, utilizing both classification tree algorithms and logistic regression. We will test these methods separately in order to assess the degree to which this enhances model stability and efficiency.

Secondary Model

The secondary model was designed to locate pro-choice voters. The choice model yields an efficiency ratio of 1.30 relative to the incidence of extremely pro-choice voters in the general population (i.e., $43/29 = 1.48$). This is higher than what would be generated by simple traditional targeting methods. As discussed above, through re-modeling we hope to increase this advantage.

Abortion Model	ST Top 5% Projection (N=255)	ST Top 5% Validation Result (N=200)	General Population (N=5000)	Traditional Voter File Targeting of Registered Voters (N=2237)
A strong supporter of a woman's right to choose (top box category=7)	58%	43%	29%	33%