



Date: April 20, 2004

To: League of Conservation Voters Education Fund

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

- LCV Education Fund SmarTargeting: The Northwest -

Executive Summary

Last year, the League of Conservation Voters Education Fund (LCVEF) asked Greenberg Quinlan Rosner Research and Stratalys Research to engage in a project designed to locate potential environmental activists in the Northwest. Using SmarTargeting techniques, we employed enhanced voter files that included information such as voting frequency, census data, and consumer data to conduct a large sample survey gauging attitudes about the environment. Using this data, we created a model that helped us identify potential environmental activists in Oregon and Washington. The modeling results were then applied back to the enhanced voter files allowing us to assign every voter in Oregon and Washington a performance score indicating how likely they are to be an environmental activist, meaning we can identify environmental activists specifically by name and address. LCVEF now has complete access to lists of potential activists around which it can design a range of different issue campaigns.

How it works:

- Each voter in the Northwest has been assigned a "predicted" performance score, or a probability that he or she is a potential environmental activist. Voters with the highest performance scores have a significantly higher chance of being an activist. For example, voters in the top decile of the environmental activist model, have a 30 percent chance of being an environmental activist. This means for every 100 people you contact, about 30 people might take some sort of environmental action. Voters in the bottom decile of the environmental activist model have a 4 percent chance of being environmental activists.

Segment	TOTAL NORTHWEST		
	Total Registered	% Enviro. Activists	# Enviro. Activists
1	574,605	30	172,381
2	574,605	20	114,921
3	574,605	16	91,936
4	574,605	14	80,445
5	574,605	15	86,191
6	574,605	12	68,953
7	574,605	12	68,953
8	574,605	10	57,461
9	574,605	8	45,968
10	574,605	4	22,984
<i>REGION</i>	5,746,055	14	810,193

- Because we are identifying potential environmental activists through a modeling process (as opposed to calling the entire population) there is error associated with these predictions, which means not everyone with the high scores will actually be activists. We have to contact a certain percentage of the top scoring voters on the list in order to reach the appropriate number; depending on how many environmental activists LCVEF wants to reach successfully.
- We also developed a secondary model to allow an even more precise voter contact strategy that allows us to select environmental activists on a more refined set of criteria. In this case, we modeled attitudes about abortion. We are able to predict which people will vote on the choice issue, which we believe is an affinity issue with the environment. In other words, there is likely overlap between people who would vote on the choice issue and those who would be active on the environment.
- Every voter in the Northwest is assigned a performance score based on the secondary model in the enhanced voter file, their predicted position on choice, allowing us to layer over additional information on our potential environmental activists.
 - To use this additional information, we could decide to refine our contact strategy by selecting every person in the top decile of the environmental activist model and overlay it with people who are in the top decile of the choice model. We could contact people who are in the top decile of both of these models – since we have their names and addresses - which would enhance LCVEF's ability to find environmental activists.
- Strategies for pulling lists from the enhanced voter files should be driven by the goals of the campaign and the resources that are available to LCVEF. This targeting system provides complete flexibility in how the LCVEF contacts potential activists whether that means by telephone, mail or canvassing.

Methodological Summary

- GQR and Stratalys Research conducted a sample survey among 5,000 registered voters in Oregon and Washington. The survey was 8 minutes long and contained a range of questions about the environment, environmental behavior and environmental issues. It also contained questions designed more generally to tap into voters' political attitudes.
- The dependent variable in the environmental model consisted of 6 variables and each one was given relative weight.
 - I am a strong environmentalist and often base my vote in elections on environmental issues (40 percent)
 - Regulation to Protect the Environment ("The laws protecting the environment are not strong enough and stronger laws need to be enacted.") (20 percent)
 - Pro-environment Consumer Behavior (Buying food that has not been chemically treated, such as organic food; Buying products made of recycled

- material whenever possible; Using public transportation, such as buses and trains.) (20 percent)
- o Membership in an environmental group (20 percent)
- The secondary model was constructed from other questions about a woman's right to choose.
 - SmarTargeting uses advanced data mining algorithms in order to create a prediction as to the likelihood an individual is an environmental activist or not. This prediction has a continuous probability value ranging between 0 and 1, with 1 representing a 100% probability that an individual is an environmental activist.^Φ The prediction score can be considered to be the weighted sum of a variety of consumer, demographic, census, and geographic variables found on the enhanced voter file. The data mining algorithms used by SmarTargeting seek out the optimal set of weights such that the predictive model approximates the definition of an environmental activist defined in the survey data (N=5000) as closely as possible.
 - We know with complete accuracy who in the survey sample is an environmental activist (according to our definition) and who is not. What we do not know is who in the larger population is an activist. Consequently, we are trying to create a model using the survey sample that can be applied to the enhanced voter files in order to derive our best guess of whether any individual is an activist. It is possible to do this because the predictors that comprise the model are found in both the survey data and the enhanced voter files. This overlap of variables allows us to apply the model to the larger population.
 - SmarTargeting utilizes a mix of different data mining methodologies that seek out relationships within the survey data in order to predict membership in the activist group with the greatest accuracy possible. These relationships are not always theoretically intuitive, and thus should not be used in order to understand how to talk to or message the activists we are targeting. The prediction score is a tactical tool to provide our best guess as to how to reach activists. However, the model itself is not specifically designed to provide message strategy.
 - Below we show a table that includes First and Second Tier Predictor variables from the models. Note that the model looks for relationships that are multiplicative and/or non-linear as well as linear. Consequently, it is difficult to make simple statements about these predictor variables of the sort, "If an individual lives in a census tract that contains a large number of individuals who are employed outside the home and commute 45 or more minutes each way, then that individual is more likely to be an environmental activist." Rather, data mining is explicitly designed to seek out more complex relationships in order to derive greater model accuracy. So, in creating a predictor score, the model assigns some number of percentage points to every individual if they are, for instance, 18-35 years of age plus some additional combination of predictor variable values (i.e., perhaps the presence of a pool in the household).

^Φ Note that while, in theory, the prediction values range between 0 and 1, it is almost always the case that actual ranges will be narrower than this. Often, the range will be approximately .05 to .80. This is simply because of the error found in any non-determinative model.

First Tier Predictor Variables
% of workforce employed in white collar professional occupations (census)
Voter State (enhanced voter file data)
% of population age 40 to 54 (census)
Voter county (census)
% of those employed outside the home who commute 45 or more minutes each way (census)
Gender (enhanced voter file data)
Pool flag (consumer data)
Age (enhanced voter file data)
Likelihood of owning/renting residence (consumer)
Presence of male 35-44 in household (consumer)
Household size (consumer)

Second Tier Predictor Variables
% of those employed outside the home who commute 60 or more minutes each way (census)
Likelihood of individual being married/single (consumer)
Number of adults in household (consumer)
% of households with unmarried partners of the same sex (census)
% of adults who are veterans
% of workforce employed in agricultural occupations (census)
Revolver minimum payment model (consumer)
Presence of male 18-34 in household (consumer)
US Postal service delivery area type (enhanced voter file data)
Presence of dog in household (consumer)
Likelihood of Internet Usage (consumer)
Financial Index (consumer)
Presence of adult age 50 or older in household (consumer)