INTRODUCING C.A.R.T. TO THE FORECASTING PROCESS

By Jeff Morrison

CART makes it easier to have an in-depth insight into different segments of data ...has advantages over regression in handling missing data and capturing nonlinearity and interactions within the data ... currently used quite successfully in the credit card industry for pre-screening credit card mailings.

The econometric literature on segmentation has historically shown that regression based forecasting models work best when applied to homogeneous groups. This makes sense for a number of reasons. Factors affecting one population may not be the same factors that impact others. In forecasting telephone lines, for example, variables describing more complex dynamics in pricing, technology, and marketing may better describe the demand for business lines than residential connections. Even within the variables themselves there may exist certain groupings that may behave differently from one another. For example, maybe the propensity to purchase consumer durables is different between high income and low income groups. Perhaps customers residing in various geographic locations behave very similarly. Although the forecaster has always had a choice to model homogeneous groups separately, not until recently has there been software available for non-statisticians to assist them in identifying similar groupings or empirical "breaks" in the data. The following discussion focuses on one such procedure called *Classification and Regression Trees* (CART).

Tree analysis along with other analytical tools such as factor analysis were originated by social scientists to provide additional insight into the data structure being studied. The use of trees in regression dates back to the AID (Automatic Interaction Detection) program developed at the Institute for Social Research, University of Michigan, by Morgan and Sonquist in the early 1960's. Largely with the help of Jerome Friedman and others in 1980, CART emerged as a practical way of interpreting data, adding another tool to the analyst's arsenal in data analysis. In the past, it has been used extensively in the medical profession to determine key factors in the risk of heart attack upon the patience's admission to the hospital. Now, with the help of new window-like software packages like S-Plus, CART business applications are seeing ever increasing use.

JEFFREY S. MORRISON

Mr. Morrison is Director of Modeling at Equifax, based in Atlanta, Ga. His work experience includes the development of new product diffusion models, quantitatively based target marketing systems, and customer satisfaction simulation models for the telephone industry. In 1992, he won the Outstanding Speaker Award at the National Telecommunication Forecasting Conference.

WHY CART?

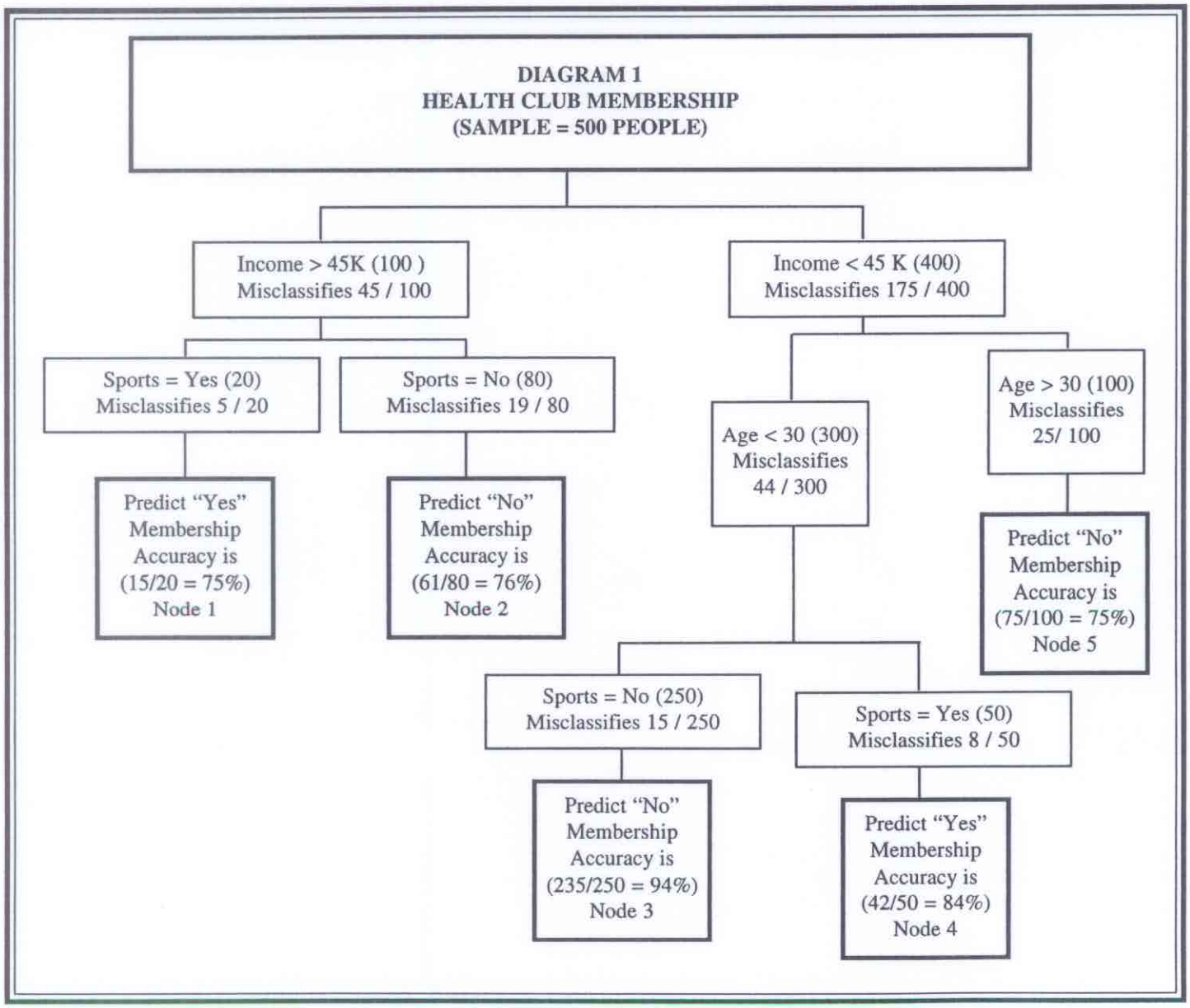
CART is a computationally intensive exploratory analysis tool which attempts to describe the structure of your data in a tree-like fashion. S-Plus, one of the more popular CART vendors, uses a measure called "deviance" to determine the tree structure. Deviance, a form of the likelihood ratio test, is used to measure the heterogeneity of the tree structure. The procedure is nonparametric, meaning that no assumptions are made as to the population's underlying distributions. This is quite different than regression analysis where statistical assumptions are essential for precision, accuracy, and interpretability. However, as in regression, CART relates a single "dependent" variable (either binary or continuous) to a set of predictors. CART's advantages over regression based approaches centers around its ability to handle "missing data" and capturing nonlinearity or interactions within the data. Although a regression model can be specified ahead of time to include interactions (income ´ price, for example), CART does it automatically with no need of user intervention.

CART is not typically applied directly at the aggregate level where most time based forecasting models are developed, i.e. predicting monthly sales over time. A better use for CART would be at the individual consumer or account level at a single snapshot in time. Since aggregate behavior at any particular point in time

should be a reflection of behavior at the consumer level, evaluating the structure of your data at its lowest level should prove insightful in segmenting your population or understanding the relationships among the variables before the time dimension is added.

Let's take a simple classification example. Suppose you wanted to predict whether or not an individual is likely to join a health club. After surveying 500 individuals, say you knew four things: (1) whether or not they participated in organized sports in high school or college, (2) their age, (3) their income level, and (4) if they were currently a member of a health club. Diagram 1 illustrates a possible tree structure in this data. Out of the 500 people surveyed, suppose 100 of them had incomes greater than $45,000 and 400 people had annual salaries less than that. Based upon information from the income criteria alone (a very naive model), the CART model, let's say, misclassified 45 out of 100 individuals with incomes greater than $45,000. That's rather poor. However, as more information is added to the CART structure (like high school sports participation), the misclassification rate drops significantly.

In our example, out of the 20 individuals who made over $45,000 and participated in sports, the misclassification rate was 5 / 20. Therefore, given these income and sports participation groupings, CART correctly classifies those individuals as owning a health club membership 75% of the time. These stopping points for CART along a branch are called "end nodes." They reflect the final prediction of whether those individuals would have purchased or not purchased a health club membership based on which group dominates the node. In the case of node 1, there were more individuals having health club memberships than not, resulting in a node 1 prediction of health club membership. For those that made over



$45,000 but did not participate in sports (80 individuals in our sample), the misclassification rate was 19 / 80. CART predicted this group (node 2) not to have a health club membership.

The right hand side branch of the tree shows a more complex structure with age playing an important role. For individuals with lower incomes, CART determined that a proper segmentation scheme appears at age 30. Segmenting the data in this way allows a greater measure of precision than otherwise. However, as you go down the tree structure on any side, note how the misclassification rate decreases. In some CART programs, you have to tell the procedure when to quit otherwise it will "overfit" the data by making each observation an end node. As in any modeling procedure, usually rules of thumb are available to keep this from happening.

CART'S PREDICTION ALGORITHM

In a regression model, the estimated coefficients are used in a forecast equation to make predictions. With CART analysis, a prediction algorithm is in the form of "if statements" for each end node. For example, node 1 & 2's prediction algorithms would be:

If Income >45000 and
if Sports = "Yes" then node = 1

If Income >45000 and
if Sports = "No" then node = 2

Individuals classified by these conditions under node 1 would be predicted to have a health club membership with a 75% accuracy rate. Individuals classified by the conditions under node 2 would have been predicted not to have a health club membership. Depending on the application, these groups could easily be collapsed into a single segment by combining the "if statements."

CART VS. REGRESSION

As illustrated earlier, using CART can yield insight into segmenting your data. Although you could develop a classification procedure using regression, how would you know where the best "breaks" are on the continuous explanatory variables? Although you could go through a laborious cross tab exercise, you could use CART to identify breaks quickly and easily. In this example, you might now design your regression model with dummy coded age variables. Assign a 1 to the variable if the individual was under 30 years old, or 0 otherwise. Then a significance test could be made using the standard t-test. The same procedure could be performed for income with a break at $45,000.

Another interesting feature of CART analysis is its ability to handle alpha-numeric data. This comes in particularly handy for geographically based segmentation schemes. For example, suppose you were doing the health club analysis and were able to obtain the state name in which the respondent resides. CART can evaluate the data's structure in either a text format such as "Florida" or coded as some numerical index like "14." CART will treat the state name in the same fashion as it would any other variable such as age, income, or even gender.

CART can also be used in regression models to add insight into what to include as explanatory factors from a large set of independent variables. Although most econometrics textbooks do not speak highly of the "stepwise selection" method, sometimes it can be justified if the economic theory behind variable selection is sound. Let's say you had 50 potential variables from a survey for individuals choosing to either purchase or not purchase a product. A stepwise procedure could narrow the field of variables down to 10. If you had used CART as a preliminary analysis tool, it could have provided insight that age and income where important factors in the decision to purchase. The stepwise selection procedure could have failed to show these variables as significant.

CART has proved useful in the credit industry for pre-screening credit card mailings and as a risk based segmentation tool. Over recent years, the search for those customers most likely to respond to credit card offers has greatly intensified. Given a short turn around time, CART provides a quick and easy way to understand what customer segments will most likely respond across different risk levels. By providing a tree like structure, the client not only can see the credit factors that are most predictive of a response, but tell at what point within those variables further segmentation may be evident. Various promotion plans can then be implemented to capture different segments of the market. An example might be to offer a "gold" package with lower interest rates to attract low risk credit seekers and a special package with higher interest charges for higher risk consumers.

When a longer turn around time is given to build a more sophisticated approach, CART still has proved useful in adding value within regression based response models. These more traditional approaches require decisions to be made as to how to divide credit data by risk of default. This may be accomplished through the use of a generic risk measure (a score ranging from 0 to 1000) derived from accounts from a variety of financial institutions. Segmentation information can be attained by placing the "response" and "non-response" data along with the generic risk score in a CART classification analysis. The CART procedure could identify the splits (or breaks such as "< 450") in the generic risk score, indicating possible segmentation schemes for regression modeling. ■