DM STAT-1 CONSULTING **BRUCE RATNER, PhD**

574 Flanders Drive North Woodmere, NY 11581 br@dmstat1.com
516.791.3544 ~ fax 516.791.5075 www.dmstat1.com

Is Not a Response-Model Tree a Response-Model Tree by Any Other Name?

Bruce Ratner, PhD

Is not a response-model tree a response-model tree by any other name? To answer the question I present two machine-learning response-model trees, one pretty popular, and one not-yet popular. [1] The *pretty* one has attained its status because nearly anyone can understand the highly accepted predictive response-model tree, regardless of one's statistical background. Yet, data analysts are not as forgiving with the not-yet popular one. The latter response-model tree requires a passing introductory training in machine learning, which data analysts typically have no such exposure. Consequently, the not-yet popular one has attained its standing, but there is a sizeable user-group of this quite predictive response-model tree.

Consider the pretty one, but hardly ever thought of as a machine learning method – CHAID. See Response CHAID Tree in Figure 1, below. The CHAID Tree can *loosely* be interpreted: The overall Response of 10% (from a population of size 1000) is explained and predicted by primarily Martial Status, and secondarily Gender and Pet Ownership.

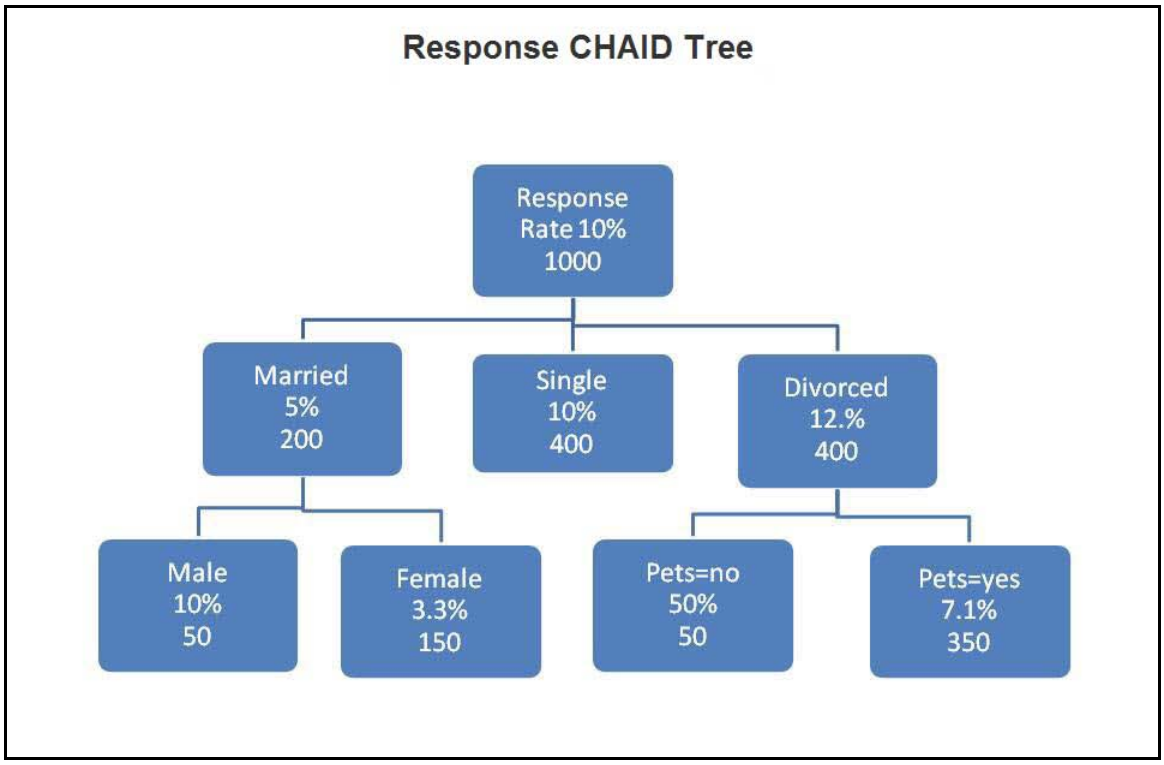

Figure 1. Response CHAID Tree

The CHAID modeling process starts out by identifying the *best* variable that predicts the response of the population (top box/node). Resultantly, the top node is split by the best variable, which yields at least two nodes. If the best variable is categorical, then the nodes are either all the original categories or a various number of combined original categories. If the best variable is continuous, then the nodes consist of at least two nodes, which nodes are defined in terms of the best variable's range: EITHER

1) A left-closed right-open internal, and a closed interval, OR
2) A various number of left-closed right-open intervals, and a closed interval. [2]

The process continues by identifying the best variable that predicts/splits each node of the *first-best* predictor. The determination of nodes is based on statistical significance tests. In Figure 1, the Married node is split by Gender; the Single node is not split because neither Gender nor Pets pass the significance test.

The splitting process is recursive: It identifies the *next-best* predictor of the previous-split nodes, and then identifies the *next-next-best* predictor of the previous-split nodes, and so on. The process stops when pre-determined stopping rules are met (e.g., node size cannot be smaller than 250 observations). In Figure 1, it is assumed that there are no additional variables to split the nodes: Males, Females, Pets=no, and Pets=yes. CHAID trees typically include, besides the first-best split of the population, two-way combination/interaction variables (like the bottom four nodes in Figure 1), and three-way interaction, four-way interactions, and sometimes five-way interactions. [3]

The not-yet popular GenIQ Model is a machine learning alternative model to the statistical ordinary least squares and logistic regression models. The GenIQ Model (using *genetic programming* as its number-crunching tool, whereas calculus is the number cruncher for the statistical models [4]) lets the data define the model – automatically data mines for new variables, performs variable selection, and then specifies the model equation – to *optimize the decile table*, to fill the upper deciles with as much profit/many responses as possible. Put differently, the GenIQ Model seeks to maximize *cum lift*, a measure of model predictiveness of identifying the upper performing individuals often displayed in a decile table.

The GenIQ Model output consists of a two-part output, a *parse* tree, and the computer code that cryptically *explains* the tree. [5] Observe the Response GenIQ Model Tree in Figure 2, and if you dare peek at the tree code in Figure 3, below. Unlike the CHAID tree, the GenIQ Model tree is hard to look at, not to mention interpret. However, the GenIQ modeling process itself is not difficult to grasp, but requires much space for a discourse, not available here. [6]

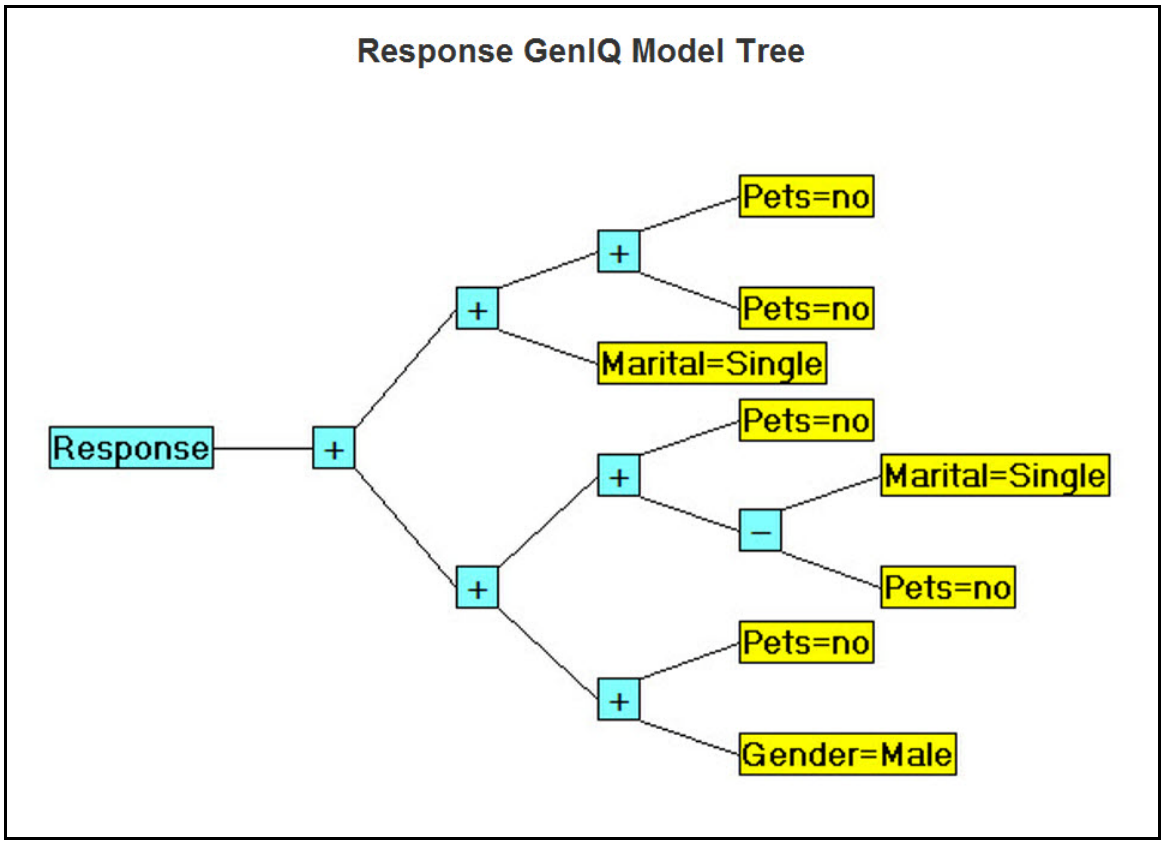

Figure 2. Response GenIQ Model Tree

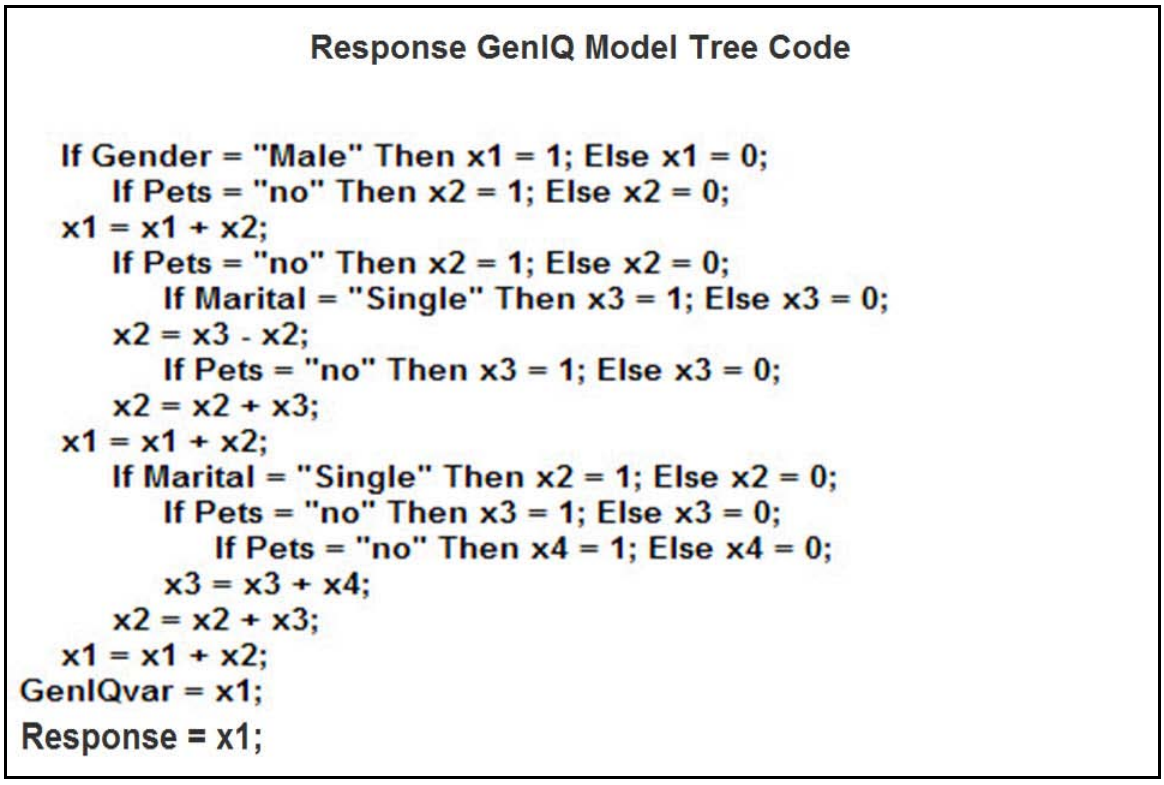

Figure 3. Response GenIQ Model Tree Code

Is not a response-model tree a response-model tree by any other name? From the brief examination above, the answer is obviously *no*. In the article proper, I deepen the discussion by comparing and contrasting the ever-popular CHAID tree, and the not-yet popular GenIQ tree.

References:

1 - Not-yet Popular Model
2 - Definitions of the intervals of nodes of a continuous predictor variable
3 - Personal Observation of Size 1: I know CHAID is widely used as an end-result model (as opposed to using CHAID as a data-mining model), but a model based on only one *main effect* predictor variable, and many, many two-, three-, and four-way interactions, which are identifying very small (i.e., unreliable) segments of the population, is not for my taste. Please keep in mind that I am not discounting CHAID. See my nine uses of CHAID beyond its original intent. [1a]
4 - What is Genetic Programming?
5 - A parse tree is a visual representation of a computer program, or a model of the type GenIQ Model produces.
6 - What is the GenIQ Model?