A Basic Introduction to CHAID

CHAID, or **Chi-square Automatic Interaction Detection**, is a Classification Tree technique that not only evaluates complex interactions among predictors, but also displays the modeling results in an easy-to-interpret tree diagram. The "trunk" of the tree represents the total modeling database. CHAID then creates a first layer of "branches" by displaying values of the strongest predictor of the dependent variable. CHAID automatically determines how to group the values of this predictor into a manageable number of categories. (E.g., we may start with ten categories of age, and CHAID might collapse these ten categories down to only four or five statistically significantly different age groups.)

CHAID then creates additional layers of branches off of each age grouping, using the strongest of the remaining predictors. It continues this branching procedure until the final branches or "twigs" of the tree have been generated. If CHAID is being used to generate a predictive market segmentation model, then these terminal branches are the final market segments.

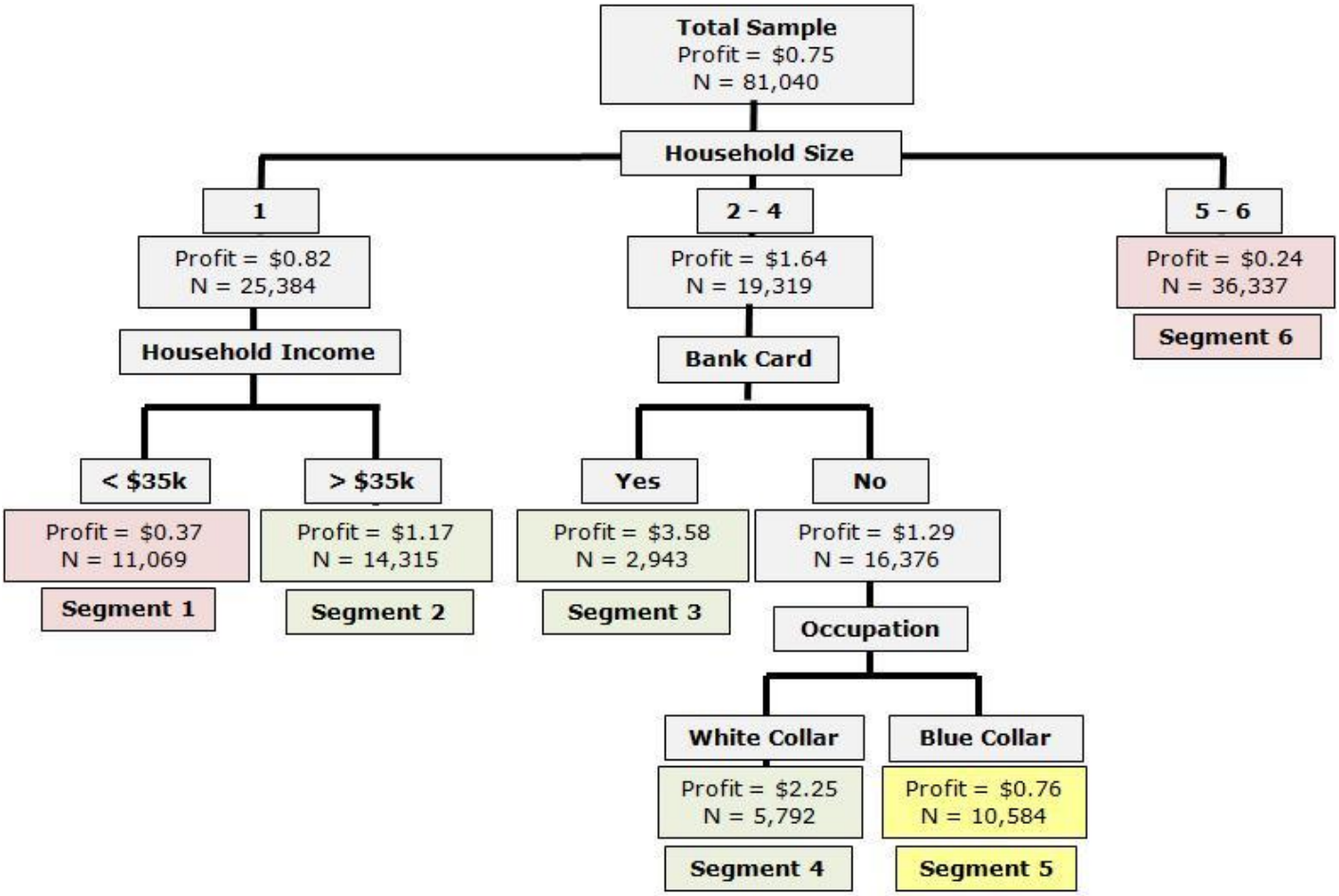
A typical CHAID model may have a dozen or so terminal segments, but sometimes we will build a model that has many more segments than that, especially if we want to identify and understand some smaller "niche" segments that may represent either a significant problem or an unusually good opportunity. The segments are depicted in the tree diagram as well as ranked in a "gains table." Statistics produced by the gains table make it easy to determine how "deep" into a file one must go to select prospects representing a given level of above-average performance (dollar value; response rate; etc.). Financial assumptions can be folded into the predictive CHAID model to generate various estimates such as ROI.

When the dependent variable we are trying to predict has only two values (e.g., mail responder vs. non-responder), we generate a **nominal CHAID model**. In such a model, we are able to see what proportion of each market segment consists of cases in the desired category of the dependent variable (e.g., mail responders). A relative performance index is generated for each segment, based on the proportion of that segment that falls into the desired category of the dependent variable.

When the dependent variable is at least ordinal (i.e., the values can be arranged in some meaningful order), then we generate an **ordinal CHAID model**. Customer dollar value is an example of an ordinal dependent variable. In an ordinal model, each segment is assigned an average value on the dependent variable (e.g., average dollar value), and this is shown in both the tree diagram and the gains table. As with the nominal CHAID model, the ordinal model can be supplemented with proprietary financial data to facilitate decision-making.

Below is a very simple, hypothetical ordinal CHAID model tree diagram for illustrative purposes only. [The data are not real.] It begins at the top with a box representing the entire modeling sample of 81,040 households (marked "Total"), to which a consumer product might be marketed via direct mail. Also included in this first box is the average profit per household generated by the initial mailing (seventy-five cents). Household size is identified by CHAID as the best predictor around which to begin segmenting the prospect market.

CHAID Tree Diagram



We can see that a household size of two to four persons returns an average profit of $1.64, which is twice the profit generated by a one-person household, and nearly seven times the profit generated by a five-to-six-person household. CHAID then shows us that if a two-to-four-person household has a bank card, the average profit jumps to $3.58. If they do not have a bank card they return an average profit of only $1.29. However, among this non-bank-card group, if the head of household's occupation is White Collar, profitability rises to $2.25.

For illustrative purposes, we have colored the above-average segments green, the average segment yellow, and the below-average segments red. Also, the segments are numbered from one to six on the tree diagram. We have used the same convention on the Gains Table, below, which displays additional useful data for the six segments displayed in the tree diagram.

CHAID Gains Table

Segment ID	Segment Count	Percent of Total	Average $ Value	Segment Index	Cum. Count	Cum. Percent	Cum. $ Value	Cum. Index
3	2,943	3.6	3.58	476	2,943	3.6	3.58	476
4	5,792	7.1	2.25	298	8,735	10.8	2.70	358
2	14,315	17.7	1.17	155	23,050	28.4	1.75	232
5	10,584	13.1	0.76	101	33,634	41.5	1.44	191
1	11,069	13.7	0.37	49	44,703	55.2	1.17	156
6	36,337	44.8	0.24	31	81,040	100.0	0.75	100

The first column of the gains table shows the segment number identifiers from the tree diagram. The second column gives the segment household counts. The third column shows what percent of the total modeling sample falls into each segment. The fourth column shows the average profit per household for each segment. The fifth column represents this profit number as a relative index, with the average for the entire modeling sample set at 100. Thus, the best segment has an index of 476, which means that it performs at a profit level of 4.76 times the average for the entire modeling sample, and more than 15 times the profitability of the worst segment.

Columns six through nine are cumulative representations of the data from columns two through five: cumulative household count, percent of modeling sample, average profit per household, and profit index. Among other things, the gains table shows us that the best three segments (segments 3, 4 and 2) represent 28.4% of the total sample, have an average profit of $1.75 per household, and are therefore 2.32 times as profitable as the average sample household.

The gains table is a handy tool for seeing what levels of expected profitability would result from going increasingly deeper into a prospect file. This is invaluable for planning direct marketing outreach programs, since it helps us determine mailing quantities, and gives us information for calculating return on investment.

If instead of dollar value, our dependent variable had simply been a dichotomous variable such as response vs. non-response to a mailing, then we would have generated a nominal CHAID analysis. In that case, instead of showing dollar values in the tree diagram and the gains table, we would have shown percent response.

CHAID is particularly useful for generating market segmentation models. In addition to its utility as a predictive modeling technique, the resulting tree diagram provides a valuable "bird's eye view" of the market structure, showing the combinations of predictors which lead to any given segment. This can be very helpful to advertising agency creatives and media planners, who want to be able to visualize and define clear market segments.

Finally, the results of a CHAID model can be used to score a master database in an easy, straightforward manner. As with other techniques (e.g., regression), new cases added to a file can be scored quickly once the basic scoring algorithm is set up.