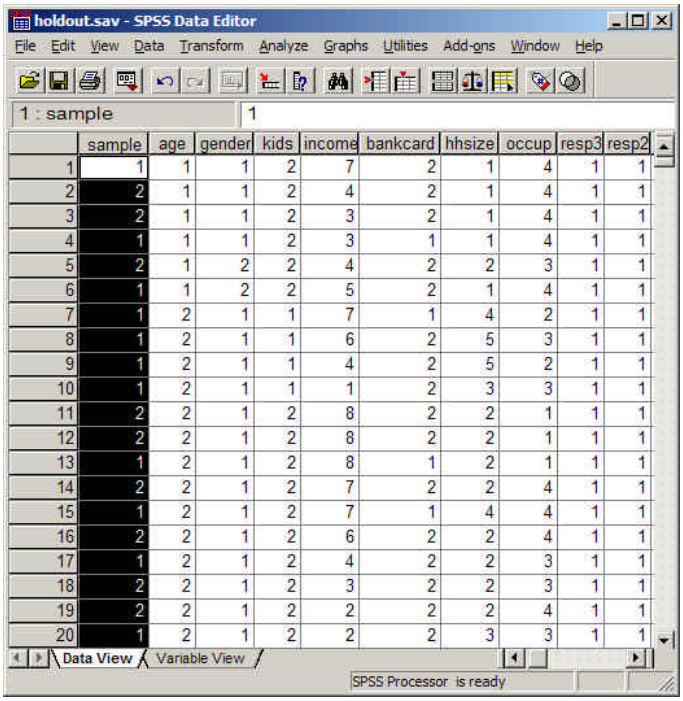
# Tutorial 3: Using SI-CHAID with a Hold-out Sample

Sometimes cases on the analysis file are randomly assigned to a 'hold-out' sample and not used in the development of the segmentation tree. Instead, such cases are reserved for the purpose of 'validating' the tree. In this tutorial we utilize the data file holdout.sav to illustrate the use of SI-CHAID in this way.

In particular, from each dependent category ('paid respondents', 'unpaid respondents' and 'non-responders') we randomly assigned each case in the 'subscrib.sav' file to one of two equally likely groups by generating the variable SAMPLE (1=test, 2 = holdout).




| | sample | age | gender | kids | income | bankcard | hhsize | occup | resp3 | resp2 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 2 | 7 | 2 | 1 | 4 | 1 | 1 |
| 2 | 2 | 1 | 1 | 2 | 4 | 2 | 1 | 4 | 1 | 1 |
| 3 | 2 | 1 | 1 | 2 | 3 | 2 | 1 | 4 | 1 | 1 |
| 4 | 1 | 1 | 1 | 2 | 3 | 1 | 1 | 4 | 1 | 1 |
| 5 | 2 | 1 | 2 | 2 | 4 | 2 | 2 | 3 | 1 | 1 |
| 6 | 1 | 1 | 2 | 2 | 5 | 2 | 1 | 4 | 1 | 1 |
| 7 | 1 | 2 | 1 | 1 | 7 | 1 | 4 | 2 | 1 | 1 |
| 8 | 1 | 2 | 1 | 1 | 6 | 2 | 5 | 3 | 1 | 1 |
| 9 | 1 | 2 | 1 | 1 | 4 | 2 | 5 | 2 | 1 | 1 |
| 10 | 1 | 2 | 1 | 1 | 1 | 2 | 3 | 3 | 1 | 1 |
| 11 | 2 | 2 | 1 | 2 | 8 | 2 | 2 | 1 | 1 | 1 |
| 12 | 2 | 2 | 1 | 2 | 8 | 2 | 2 | 1 | 1 | 1 |
| 13 | 1 | 2 | 1 | 2 | 8 | 1 | 2 | 1 | 1 | 1 |
| 14 | 2 | 2 | 1 | 2 | 7 | 2 | 2 | 4 | 1 | 1 |
| 15 | 1 | 2 | 1 | 2 | 7 | 1 | 4 | 4 | 1 | 1 |
| 16 | 2 | 2 | 1 | 2 | 6 | 2 | 2 | 4 | 1 | 1 |
| 17 | 1 | 2 | 1 | 2 | 4 | 2 | 2 | 3 | 1 | 1 |
| 18 | 2 | 2 | 1 | 2 | 3 | 2 | 2 | 3 | 1 | 1 |
| 19 | 2 | 2 | 1 | 2 | 2 | 2 | 2 | 4 | 1 | 1 |
| 20 | 1 | 2 | 1 | 2 | 2 | 2 | 3 | 3 | 1 | 1 |

*Figure 38. Holdout.sav file*

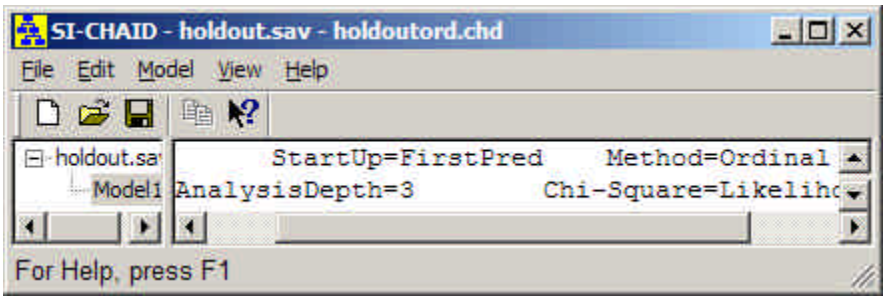
In this tutorial we will use this data file to grow a segmentation tree on the test file and see how well it validates on the holdout sample. This will be accomplished using the following steps:

- Use the 'First predictor' option to force the variable SAMPLE (test vs. holdout) to yield the first split
- Use the 'auto' option to grow the tree only on the SAMPLE = test group
- Save the resulting tree
- Apply the saved tree to the SAMPLE = 'holdout' group
- Compare gains-charts for the test and holdout samples

➢ From the Define program, select File Open 'holdout.chd'

Your display should now look like Figure 39. Note that the options shown in the Contents Pane indicate that the tree will be grown using the file 'holdout.sav' with the First Predictor option and the Ordinal method.

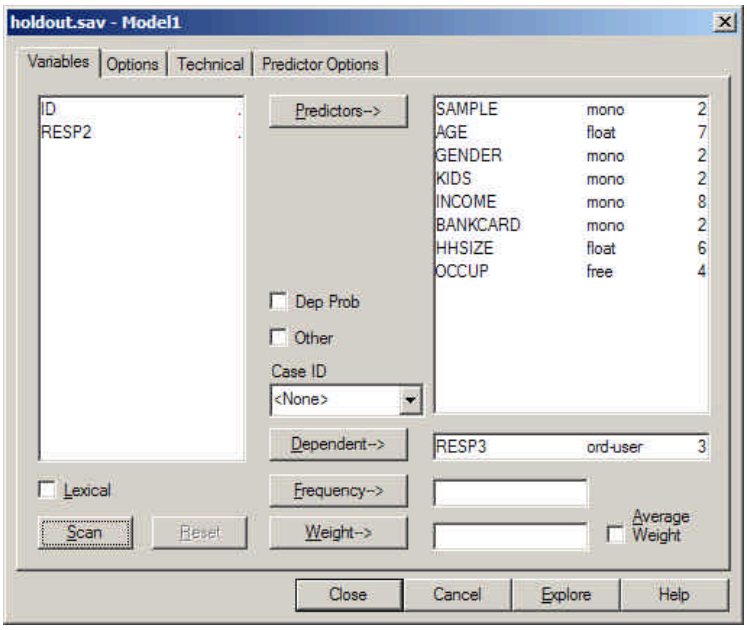
*Figure 39. Holdout.sav in Chaid Define*




To open the analysis dialog box:

- From the Model menu select 'Edit' (or double click on 'Model1')
- Click Scan

*Figure 40. Analysis Dialog Box for Holdout.sav*




Note that the dependent, predictor variables and scale types are identical to that used in the ordinal model developed in Tutorial #2, except that the new variable SAMPLE is used as the first predictor.

- Click 'Options' to open the Options tab

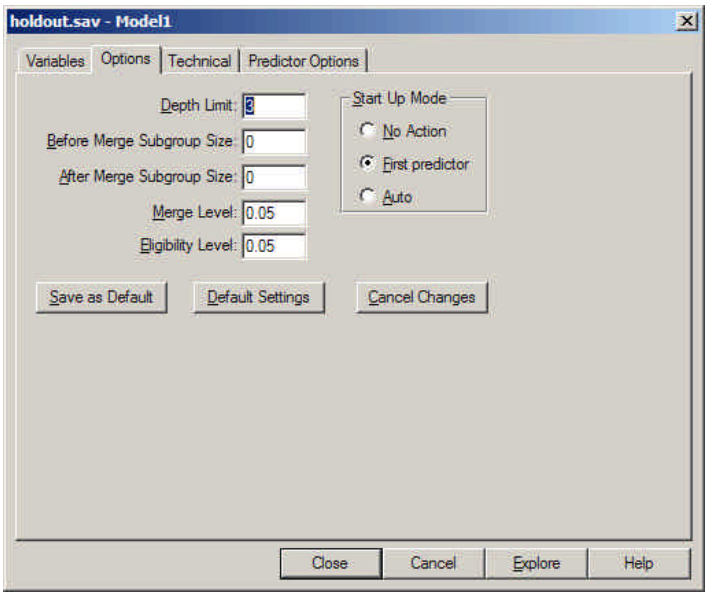



*Figure 41. Options Tab for Holdout.sav*

The 'First Predictor' option means that the categories of the first predictor variable SAMPLE will be used to define the initial CHAID split. This is indicated in the Start-Up Mode box.

- Click Explore
- When prompted, enter the file name 'holdout.chd'
- Select Yes, to replace the current file of the same name

The Explore program opens and grows the tree to one level, using the 2 categories of SAMPLE as shown below.

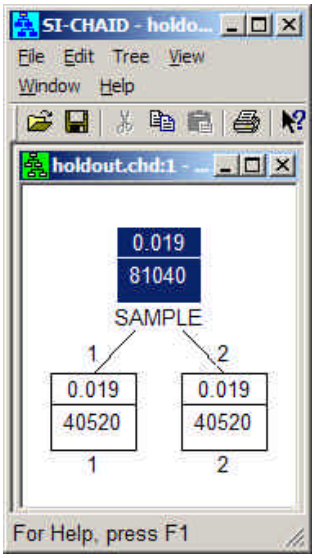



*Figure 42. Tree Diagram for SAMPLE*

The contents of the nodes shows that both the SAMPLE = 1 (test group) and SAMPLE = 2 (holdout group) consist of exactly half of the cases (N=40,520), each having an average profit of $.019 per case.

To grow the tree within the test sample,

- Click on node 1
- From the Tree menu, select auto

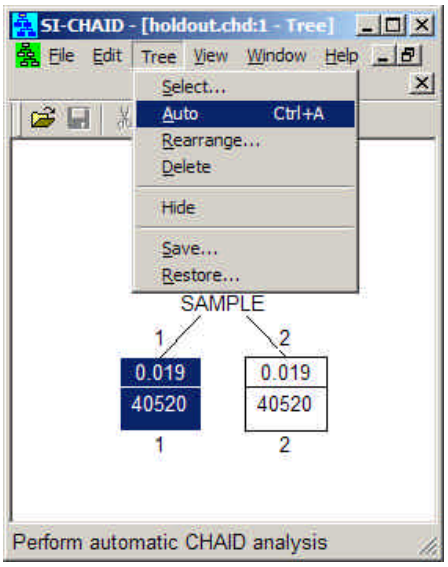



*Figure 43. Selecting Auto from the Tree menu*

The resulting tree consists of 5 segments, numbered 1-5. Segment #2 shows the highest profit ($.467), followed by segment # 4 ($.237), segment #3 ($.102), segment #1 ($.043) and segment #5 (-$.061).

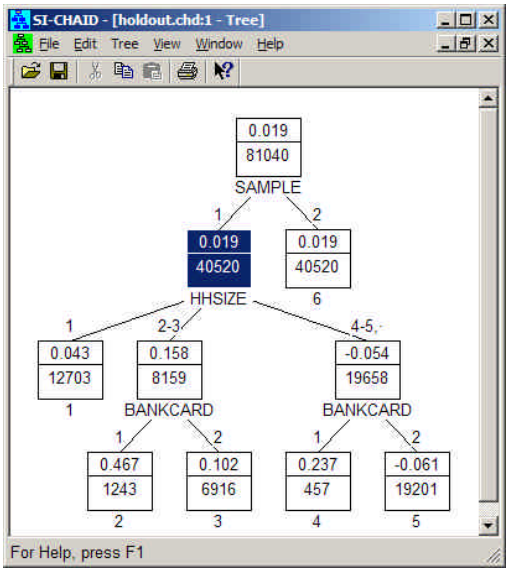



*Figure 44. 5 segment Tree Diagram*

One way to apply this tree to the holdout sample is to

- Select Edit>Copy
- Click on node #6
- Select Edit>Paste

An alternative approach is to save the tree to a file and then restore it to the holdout sample

To save the tree in Figure 44 corresponding to SAMPLE=1,

- from the Tree menu, select Save
- when prompted for a file name, enter '5segments.ctf
- Click Save

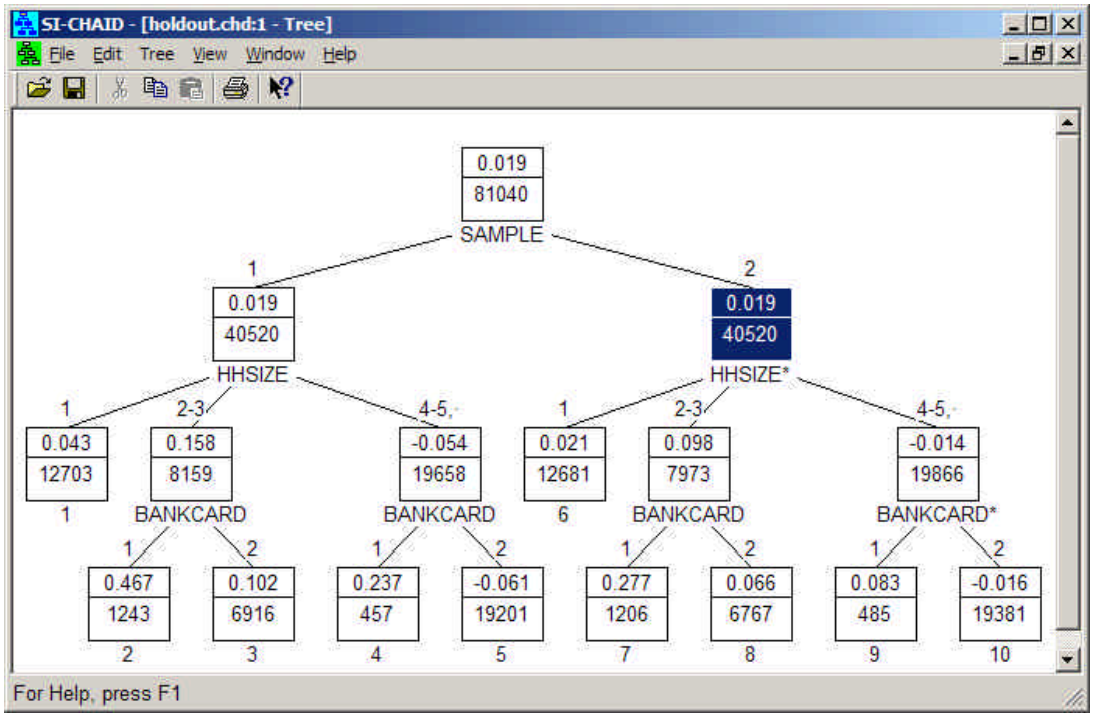
The CHAID tree file '5segments.ctf' is saved

To apply this tree to the holdout sample,

- click on node #6
- from the Tree menu, select Restore
- When prompted for a file, select '5segments.ctf'
- Click Open

Regardless of which way you chose to apply the tree to the holdout sample, your display will now look like this:

*Figure 45. Tree applied to the holdout sample*



To compare gains charts for the test and hold-out samples:

- First, click on the Parent node associated with SAMPLE =1.
- From the Windows menu, select 'New Gains'

The following Detail view of the Gains Chart appears:

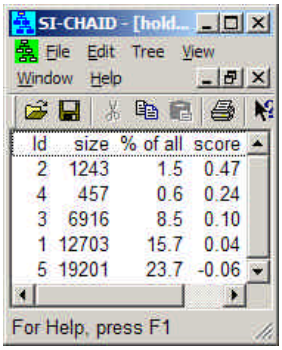



*Figure 46. Gains Chart of the Holdout Sample*

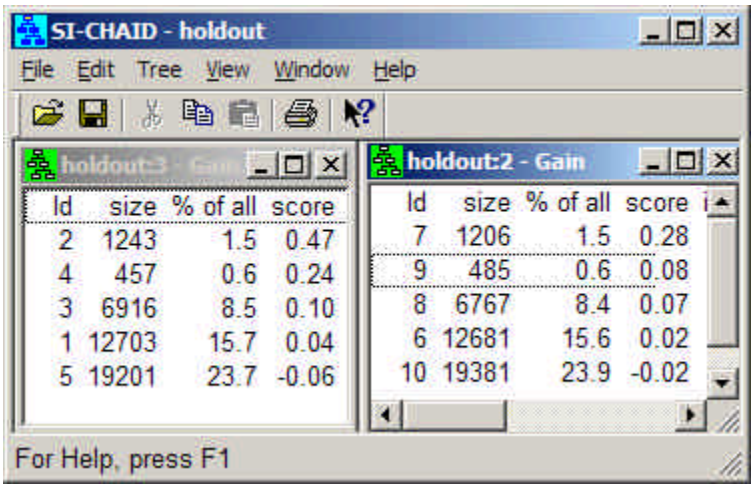
The segments are sorted from best to worst. The first segment corresponds to node #2, with a score of $0.47. (Note that in the Tree Diagram, this is displayed to an additional decimal place -- 0.467. To fix this gains chart so it will not change when we make the node SAMPLE = 2 the current node:

- Right click on the gains chart to retrieve the Gains Items control panel
- Select Fixed

- Now, click on the Parent node associated with SAMPLE =2.
- From the Windows menu, select 'New Gains'
- Right-click on the new Gains Chart
- Select Fixed

These gains charts may be used to validate the tree.

- Rearrange the 2 Gains Charts so they appear side by side:




| Id | size | % of all | score |
|---|---|---|---|
| 2 | 1243 | 1.5 | 0.47 |
| 4 | 457 | 0.6 | 0.24 |
| 3 | 6916 | 8.5 | 0.10 |
| 1 | 12703 | 15.7 | 0.04 |
| 5 | 19201 | 23.7 | -0.06 |

| Id | size | % of all | score |
|---|---|---|---|
| 7 | 1206 | 1.5 | 0.28 |
| 9 | 485 | 0.6 | 0.08 |
| 8 | 6767 | 8.4 | 0.07 |
| 6 | 12681 | 15.6 | 0.02 |
| 10 | 19381 | 23.9 | -0.02 |

*Figure 47. The two gains charts side-by-side*

Notice first, that the rank ordering of the segments in the test sample is found to validate perfectly the holdout sample. Thus, the best group to target would be segment #2 (which corresponds to node #7 in the holdout sample), next segment #4 (node #9 in the holdout sample), etc.

Note that the gain from mailing to the best segment is estimated to be $.28 (per mail piece) using the holdout cases, which is lower than the gain of $.47 estimated using the test cases. Similarly, the loss estimated associated with mailing to the worst segment (segment #5) is estimated to be less extreme using the holdout cases (-$.02 vs. -$.06). Such 'regression to the mean' is a natural phenomenon, which can be expected to occur in test validation exercises such as this.

The estimates obtained from the holdout sample are unbiased estimates of what would be likely to occur in a rollout. The extent of the 'regression to the mean' falloff may be interpreted as a measure of the amount of 'overfitting' that is present in the original model developed on the test sample. The expected amount of falloff is in part a function of the sample size. Thus, a CHAID tree developed on all n=81,040 cases as was done in Tutorial #3, would be expected to result in less falloff than this CHAID tree. That is why many researchers do not use a holdout sample when estimating CHAID or other statistical models.