Securities Brokerage CHAID Model

Note: if you are not highly familiar with **CHAID** segmentation modeling techniques, we suggest that you first visit our **Introduction to CHAID** page, and then return here to continue reading the brokerage case study.

Introduction

This case study documents a predictive market segmentation model designed to identify and profile high-value brokerage customer segments as targets for special marketing communications efforts. The dependent variable for this ordinal CHAID model is brokerage account commission dollars during the past 12 months. Predictors include proprietary client data (various account status and trading behavior variables) as well as syndicated demographic and lifestyle variables.

We begin by splitting the client's entire customer file into a modeling sample and a validation sample. (Once the model is built using the modeling sample, we apply it to the validation sample to see how well it works on a sample other than the one on which it was built).

CHAID Segmentation Model

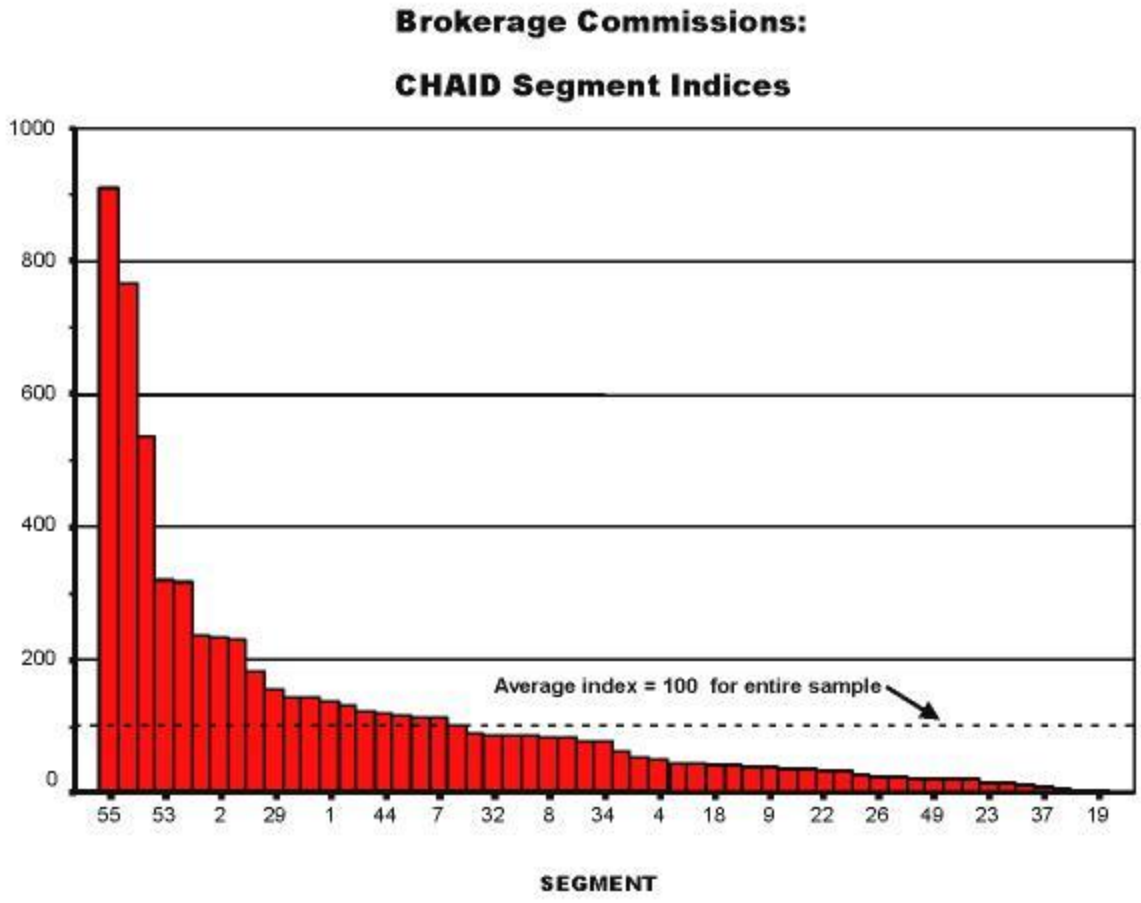
CHAID (Chi-square Automatic Interaction Detection) is a predictive, tree-based segmentation technique that uses a dependent or criterion variable as the seed for the segmentation. The tree is built by splitting an initial predictor variable into categories or ranges of values such that each category or range of values represents a statistically significant difference on the dependent variable. Additional predictors may be added under each category or value range of the first predictor, adding more branches to the segmentation model. A stopping algorithm determines when the model is complete, and the final "twigs" of the segmentation tree represent the final

segments.

Such a tree can be developed in automatic tree-generation mode, or it can be handcrafted by manually selecting predictors to use at various levels of the tree. Often a tree is first generated automatically; then, based on the specific business objectives of the analysis, as well as knowledge of the product/service category being analyzed, the tree is manually revised.

In our example, the resulting CHAID model has 55 segments. [NOTE: while a typical CHAID segmentation model might optimally contain perhaps 12-15 segments, occasionally it is useful to develop a more extensive, more granular segmentation, which can sometimes help us identify and understand important "niche" segments that represent either a significant problem or an unusually lucrative opportunity. A more extensive segmentation model also allows us to take finer cuts at a marketing database. This can yield targeting precision that is functionally equivalent to that achieved by regression modeling, while also giving us a market segmentation that regression techniques are unable to provide.]

For reasons of confidentiality, we will not display the segmentation tree diagram here. However, the results are summarized in the following comb chart, showing the segment indexes (indexes of average dollar value), and in the associated Gains table, farther down.



This summary comb chart is a quick way to confirm that the CHAID segmentation model discriminates degrees of customer value quite well. The top segment has an index of just over 900, reflecting an average annual commission value of just over $1,000. In contrast, the bottom segment has an index of just two, reflecting an average annual commission of less than three dollars.

Next is the gains table, which provides quantitative detail useful for financial and marketing planning. In the gains table, we have highlighted the top 20% of the file in blue, the remaining above-average segments in green, the bottom 20% of the file in bright red the remaining below-average segments in dark red. Among other things, we can see that the top 20% of the file is worth an average of about $334 per account, which is nearly three times the average account value for the entire sample.

CHAID Gains table: Average Annual Brokerage Commission Dollars

Basic Segment Statistics					Cumulative Statistics			
Seg. Number	Seg. Size	Percent of all	Avg. $ value	Index of avg. $ value	Cum. size	Cum. % of all	Cum. avg. $ value	Cum. index of avg. $ value
55	545	1.4	1025	907	545	1.4	1025	907
43	417	1.1	862	763	962	2.5	954	845
54	469	1.2	604	534	1,431	3.7	839	743
53	449	1.2	359	318	1,880	4.9	725	641
42	617	1.6	354	313	2,497	6.5	633	560
51	467	1.2	263	233	2,964	7.7	575	509
2	382	1	259	230	3,346	8.7	539	477
40	556	1.4	257	228	3,902	10.2	499	441
46	347	0.9	204	181	4,249	11.1	475	420
29	927	2.4	174	154	5,176	13.5	421	372
11	658	1.7	160	142	5,834	15.2	391	346
35	484	1.3	159	141	6,318	16.4	373	331
1	924	2.4	153	135	7,242	18.8	345	306
16	439	1.1	147	130	7,681	20	334	296
52	866	2.3	134	119	8,547	22.2	314	278
44	476	1.2	132	117	9,023	23.5	304	269
39	360	0.9	129	115	9,383	24.4	297	263
3	966	2.5	127	112	10,349	26.9	282	249
7	807	2.1	125	111	11,156	29	270	239

38	725	1.9	111	98	11,881	30.9	261	231
14	1,081	2.8	97.13	86	12,962	33.7	247	219
32	1,123	2.9	96.27	85	14,085	36.7	235	208
28	583	1.5	94.31	83	14,668	38.2	229	203
41	339	0.9	93.26	83	15,007	39.1	226	200
8	842	2.2	93.14	82	15,849	41.3	219	194
48	374	1	90.56	80	16,223	42.2	216	191
25	760	2	84.94	75	16,983	44.2	210	186
34	627	1.6	84.68	75	17,610	45.8	206	182
6	920	2.4	66.48	59	18,530	48.2	199	176
36	1,363	3.5	57.97	51	19,893	51.8	189	168
4	384	1	53.51	47	20,277	52.8	187	165
12	2,314	6	50.36	45	22,591	58.8	173	153
21	676	1.8	50.28	44	23,267	60.6	169	150
18	2,151	5.6	46.43	41	25,418	66.2	159	141
13	498	1.3	45.85	41	25,916	67.5	157	139
24	674	1.8	45.04	40	26,590	69.2	154	136
9	906	2.4	43.81	39	27,496	71.6	150	133
33	605	1.6	40.88	36	28,101	73.1	148	131
47	386	1	39.57	35	28,487	74.1	146	130
22	491	1.3	37.76	33	28,978	75.4	145	128
45	458	1.2	37.56	33	29,436	76.6	143	126
30	391	1	31.35	28	29,827	77.6	141	125
26	562	1.5	27.84	25	30,389	79.1	139	123
10	763	2	27.59	24	31,152	81.1	137	121

15	305	0.8	24.64	22	31,457	81.9	135	120
49	617	1.6	24.54	22	32,074	83.5	133	118
5	321	0.8	23.78	21	32,395	84.3	132	117
31	432	1.1	23.44	21	32,827	85.4	131	116
23	336	0.9	15.95	14	33,163	86.3	130	115
17	632	1.6	15.66	14	33,795	88	128	113
20	396	1	12.36	11	34,191	89	126	112
37	1,071	2.8	9.2	8	35,262	91.8	123	109
27	2,203	5.7	6.39	6	37,465	97.5	116	102
50	578	1.5	3.27	3	38,043	99	114	101
19	377	1	2.33	2	38,420	100	113	100

Additional Analyses

We can use the data in the gains table to perform various financial calculations. For example, by multiplying the size of one or more segments by the average segment dollar value, we get a total value. Using this information, we can better plan our communications/promotion budget. Similar calculations performed on the under-performing segments provide information about potential cost savings achieved by reducing marketing efforts directed at these under-performers.

We can perform additional diagnostics on the most lucrative segments to identify those customers who are below average for their segment. This means that they have the same or similar characteristics as their better-performing cohorts on the model's predictive variables, but are not providing as high a level of commission dollars. By using the demographic, lifestyle and trading behavior variables to define them, we can develop marketing and advertising communications strategies tailored to them, the goal being to convince them to trade at levels similar to their segment's more lucrative counterparts.

There are many other decisions which the gains table and the segmentation rules can help us make. For example, we might wish to conduct some market research among customers in under-performing segments, or among under-performing customers in the better segments. We can use the variables that comprise the segment definitions to help us identify possible issues

and question areas to include in the survey.

CHAID Model Validation

Before we try to apply a CHAID model, we perform a validation against a holdout sample, to confirm that it is a good model. In this example, our database was split randomly into equal-sized modeling and validation samples. We can reconstruct the CHAID segmentation model on the validation sample, and examine the results. For example, when we perform correlation analyses on the segments from the two samples, we obtain a correlation of approximately 0.98 for the segment sizes and approximately .97 for the segments' average dollar value, indicating a very high degree of correspondence.