Identifying the best customers: Descriptive, predictive and look-alike profiling

Received (in revised form): 8th April, 1999

Dr Bruce Ratner

is President of DM STAT-1 CONSULTING, the leading firm for analysis and modelling in the DM industry, specialising in statistical methods and knowledge discovery and data-mining tools, in the areas of banking, insurance, finance, retail, telecommunications, business-to-business, and catalogue marketing. Bruce is the author of the *DM STAT-1 Newsletter* on the Web at www.dmstat1.com.

Abstract Direct marketers typically attempt to improve the effectiveness of their campaigns by targeting their best customers. In the author's experience, many direct marketers are unaware that typical target methods develop a descriptive profile of their target customer — an approach that often results in less than successful campaigns. The purpose of this paper is to illustrate the inadequacy of the descriptive approach and to demonstrate the benefits of the correct predictive profiling approach. The predictive profiling approach is explained, and then the approach to look-alike profiling is expanded.

DEFINITIONS

It is helpful to have a general definition of each of the three concepts discussed in this paper. Descriptive profiles report the characteristics of a group of individuals. These profiles do not allow for drawing inferences about the group. The value of a descriptive profile lies in its definition of the salient characteristics of a target group, which are used to develop an effective marketing strategy.

Predictive profiles report the characteristics of a group of individuals. These profiles do allow for drawing inferences about a specific behaviour, such as response. The value of a predictive profile lies in its predictions of the behaviour of individuals in a target group, which are used in producing a list of likely respondents to a direct marketing campaign.

A look-alike profile is a predictive profile based on a group of individuals who look like the individuals in a target group. When resources do not allow for the gathering of information on a target group, a predictive profile built on a surrogate or 'look-alike' group provides a viable approach for predicting the behaviour of the individuals in the target group.

ILLUSTRATION OF A FLAWED TARGETING EFFORT

Consider a hypothetical test mailing to a sample of 1,000 individuals conducted by Cell-Talk, a cellular phone carrier promoting a new bundle of phone features. Three hundred individuals responded, yielding a 30 per cent response rate. (The offer also included

Bruce Ratner
DM STAT-1 Consulting, 574
Flanders Drive, North
Woodmere, NY 11581,
USA.
Tel: +1 516 791 3544;
Fax: 1 516 791 5075;
e-mail: br@dmstat1.com

Table 1: Respondents and non-respondents' profiles, response rates by gender

Gender	Respondents		Non-respondents		Response rate %
	Count	%	Count	%	
Female	30	10	70	10	30
Male	270	90	630	90	30
Total	300	100	700	100	

Table 2: Respondents and non-respondents' profiles, response rates by own_cell

OWN_CELL	Respondents		Non-respondents		Response rate %
	Count	%	Count	%	
Yes	165	55	385	55	30
No	135	45	315	45	30
Total	300	100	700	100	

the purchase of a cellular phone for individuals who do not have one, but now want one because of the attractive offer.) Cell-Talk analysed the respondents and profiled them by using the following variables: GENDER and OWN_CELL (current cellular phone ownership). Ninety per cent of the 300 respondents are male, and 55 per cent already own a cellular phone. Cell-Talk concluded the typical respondent is a male and owns a cellular phone. See Tables 1 and 2.

Cell-Talk plans to target the next 'features' campaign to males and to owners of cellular phones. The effort is sure to fail. The reason for the poor prediction is that the profile of their best customers (respondents) is descriptive, not predictive. That is, the descriptive respondent profile describes respondents without regard to responsiveness,¹ and therefore the profile does not imply that the best customers as defined are responsive.

Using a descriptive profile for predictive targeting draws a false implication of the descriptive profile. In the example, the descriptive profile of '90 per cent of the respondents are males' does not imply that 90 per cent of males are respondents, or even that males are more likely to respond.²

Additionally, '55 per cent of the respondents who own cellular phones' does not imply that 55 per cent of cellular phone owners are respondents, or even that cellular phone owners are more likely to respond.

The value of a descriptive profile lies in its definition of the salient characteristics of the best customers, which are used to develop an effective marketing strategy. In the example, knowing that the target customer is a male and owns a cellular phone, the campaign offer is positioned with a man wearing a cellular phone on his belt, instead of a woman reaching for a cellular phone in her purse. Accordingly, a descriptive profile tells the marketer how to talk to the target audience, and, as will be seen in the next section, a predictive profile helps the marketer find the target audience.

WELL-DEFINED TARGETING EFFORT

A predictive profile describes respondents with regard to responsiveness, ie in terms of variables that discriminate between respondents and non-respondents. Effectively, the discriminating or predictive variables produce varied

Figure 1: GENDER Tree

response rates, and imply an expectation of responsiveness. To clarify this, consider the response rates for GENDER in Table 1. The response rates for both males and females are 30 per cent. Accordingly, GENDER does not discriminate between respondents and non-respondents (in terms of responsiveness). There are similar results for OWN_CELL. See Table 2.

Hence, GENDER and OWN_CELL have no value as predictive profiles. Cell-Talk's targeting of males and current cellular phone owners is expected to generate the average (sample) response rate of 30 per cent. In other words, this profile in a targeting effort will not produce more respondents than will a random sample.

A new variable, CHILDREN, is now introduced which is hoped to have predictive value. CHILDREN is defined as 'yes' if individual belongs to a household with children; 'no' if individual does not belong to a household with children. Instead of discussing CHILDREN using a tabular display (such as in Tables 1 and 2), a user-friendly visual display known as Trees³ is preferred.

Response rates are best illustrated by use of Tree displays. Figures 1 and 2 review the GENDER and OWN_CELL variables in a Tree display. From this point on, only the Tree in this discussion is referred to, underscoring the utility of

a Tree as a profiler and reducing the details of Tree building to non-technical summaries.

The tree for GENDER in Figure 1 is read as follows:

- the top box indicates that for the sample of 1,000 individuals there are 300 respondents, and 700 non-respondents. The response rate is 30 per cent and non-response rate is 70 per cent
- the left box represents 100 females consisting of 30 respondents and 70 non-respondents. The response rate among the 100 females is 30 per cent
- the right box represents 900 males consisting of 270 respondents and 630 non-respondents. The response rate among the 900 males is 30 per cent.

The tree for OWN_CELL in Figure 2 is read as follows:

- the top box indicates that for the sample of 1,000 individuals there are 300 respondents, and 700 non-respondents. The response rate is 30 per cent and non-response rate is 70 per cent
- the left box represents 550 individuals who do own a cell phone. The response rate among these individuals is 30 per cent
- the right box represents 450

Figure 2: OWN_CELL Tree

Figure 3: CHILDREN Tree

individuals who do not own a cell phone. The response rate among these individuals is 30 per cent.

The new variable CHILDREN is defined as presence of children in the household (yes/no). The CHILDREN tree in Figure 3 is read as follows:

- the top box indicates that for the sample of 1,000 individuals there are 300 respondents, and 700 non-respondents. The response rate is 30 per cent and non-response rate is 70 per cent
- the left box represents 545 individuals belonging to households with children. The response rate among these individuals is 45.9 per cent
- the right box represents 455 individuals belonging to households with no children. The response rate

among these individuals is 11 per cent.

CHILDREN has value as a predictive profile, as it produces varied response rates. If Cell-Talk targets individuals belonging to households with children, the expected response rate is 45.9 per cent. This represents an increase in response over the average (sample) response by a 'lift' of 153 (lift equals profile response rate 45.9 per cent divided by sample response rate 30 per cent times 100). That is, the predictive profile is expected to produce 1.53 times more respondents than expected from a random sample.

PREDICTIVE PROFILES

Using additional variables, a single variable tree can be grown into a full

tree with many interesting and complex predictive profiles, although actual building of a full tree is beyond the scope of this paper. Suffice to say, a tree is grown to create end-node profiles or segments with the greatest variation in response rates across all segments. A tree has value as a set of predictive profiles to the extent (1) the number of segments with response rates greater than the sample response rate is large and (2) the corresponding segment lifts (ie segment response rate/sample response rate) are large.⁴

Consider the full tree defined by GENDER, OWN_CELL and CHILDREN in Figure 4. The tree is read as follows:

- the top box indicates that for the sample of 1,000 individuals there are 300 respondents, and 700 non-respondents. The response rate is 30 per cent and non-response rate is 70 per cent
- the end-node segments are referenced from left to right: #1 through #7
- segment #1 represents 30 females belonging to households with children and own a cellular phone. The response rate among these individuals is 50 per cent
- segment #2 represents 15 females belonging to households with children and do not own a cellular phone. The response rate among these individuals is 100 per cent
- segment #3 represents 300 males belonging to households with children and own a cellular phone. The response rate among these individuals is 40 per cent
- segment #4 represents 200 males belonging to households with children and do not own a cellular phone. The response rate among these individuals is 50 per cent
- segment #5 represents 55 females belonging to households with no

children. The response rate among these individuals is 0 per cent

- segment #6 represents 200 males belonging to households with no children and own a cellular phone. The response rate among these individuals is 15 per cent
- segment #7 represents 200 males belonging to households with no children and do not own a cellular phone. The response rate among these individuals is 10 per cent.

Four segments have response rates greater than the sample response rate of 30 per cent. A targeted effort to any one of these segments is expected to produce an increase in response over a random selection by a segment lift between 333 to 133. However, the top segment, in terms of response rate, is relatively small. Specifically, segment #2 consists of 1.5 per cent (15/1000) of the population. If the population itself is large, then targeting this segment will yield a campaign effort with a substantial number of pieces. For example, if the population consists of 1.5m individuals, then targeting segment #2 yields a campaign of size 22,500 with a whopping expected 100 per cent response rate⁵ (see Table 3).

If the population is not large, then several of the best-responding segments should be combined to yield a campaign with a large number of pieces. Here, the recommendation is for a mailing consisting of the top three segments, which would account for a depth of 24.5 per cent of the population. For example, if population is 500,000 individuals, then targeting segments #2, #1 and #3 yields a campaign of size 122,500 with an expected 45.9 per cent response rate (see Table 3).

Specifically, Table 3 tells Cell-Talk that it can expect:

Figure 4: Full tree defined by GENDER, OWN_CELL and CHILDREN

Table 3: Gains chart for tree defined by gender, own_cell, and children
 *(Segments are ranked by response rates)

Segment*	Size of group	Number of responses	Cumulative responses	Response rate %	Cumulative rate %	Cumulative lift
#2 — OWN_CELL, no GENDER, female, CHILDREN, yes	15	15	15	100.0	100.0	333
#1 — OWN_CELL, yes GENDER, female, CHILDREN, yes	30	15	30	50.0	66.7	222
#3 — OWN_CELL, no GENDER, male, CHILDREN, yes	200	100	130	50.0	53.1	177
#4 — OWN_CELL, yes GENDER, male, CHILDREN, yes	300	120	250	40.0	45.9	153
#6 — OWN_CELL, yes GENDER, male, CHILDREN, no	200	30	280	15.0	37.6	125
#7 — OWN_CELL, no GENDER, male, CHILDREN, no	200	20	300	10.0	31.7	106
#5 — GENDER, female CHILDREN, no	55	0	300	0.0	30.0	100
	1000	300		30.0		

- cumulative lift of 333 by targeting the top segment, which accounts for only 1.5 per cent of the population
- cumulative lift of 222 by targeting the top two segments, which account for only 4.5 per cent ($\{15 + 30\}/1000$) of the population
- cumulative lift of 177 by targeting the top three segments, which account for 24.5 per cent ($\{15 + 30 + 200\}/1000$) of the population
- cumulative lift of 153 by targeting the top four segments, which account for 54.5 per cent ($\{15 + 30 + 200 + 300\}/1000$) of the population. Unless the population is small, a campaign targeted at the top four segments may be cost prohibitive.

CONTINUOUS TREES

So far, the profiling uses only categorical variables, that is, variables that assume two or more discrete values. Fortunately,

trees can accommodate continuous variables, or variables that assume many numerical values, which allows for developing profiles with both categorical and continuous variables.

Consider INCOME, a new variable. The INCOME tree in Figure 5 is read as follows:

- Tree notation: Trees for a continuous variable denote the continuous values in ranges: a closed interval, or a left-closed/right-open interval. The former is denoted by $[a, b]$ indicating all values between and including a and b . The latter is denoted by $[a, b)$ indicating all values greater than or equal to a , and less than b
- the top box indicates that for the sample of 1,000 individuals there are 300 respondents, and 700 non-respondents. The response rate is 30 per cent and non-response rate is 70 per cent
- segment #1 represents 300 individuals with income in the interval $[\$23,000,$

Figure 5: INCOME Tree

- segment #1 represents 700 individuals with income in the interval [$\$23,000, \$75,000$). The response rate among these individuals is 30 per cent
- segment #2 represents 200 individuals with income in the interval [$\$75,000, \$157,000$). The response rate among these individuals is 10 per cent
- segment #3 represents 500 individuals with income in the interval [$\$157,000, \$250,000$]. The response rate among these individuals is 30 per cent.

The determination of the number of nodes and the range of an interval is based on a computer-intensive iterative process, which tests all possible numbers and ranges. That is, a tree is grown to create nodes/segments with the greatest variation in response rates across all segments.

The full tree with the variables, GENDER, OWN_CELL, CHILDREN and INCOME is displayed in Figure 6. Cell-Talk can expect a cumulative lift of 177, which accounts for 24.5 per cent of the population, by targeting the top three segments (see Table 4).

It is interesting to compare this tree, which includes INCOME, to the tree without INCOME in Figure 4. Based on Tables 3 and 4, these two trees have the same performance statistics, at least for the top three segments, a cumulative lift of 177, accounting for 24.5 per cent of the population.

This raises some interesting questions. Does INCOME add any significant predictive power? In other words, how important is INCOME? Which tree is better? Which set of variables is best? There is a simple answer to these questions (and many more tree-related questions): an analyst can grow many equivalent trees to explain the same response behaviour. The tree that suits the analyst is the best (at least for that analyst). Again, detailed answers to these questions (and more) are beyond the scope of this paper.

LOOK-ALIKE PROFILING

In the Cell-Talk illustration, Cell-Talk requires predictive profiles to increase response to its campaign for a new bundle of features. They conduct a test mailing to obtain a group of their best customers, respondents of the new bundle offer, on which to develop the profiles.

Now, consider that Cell-Talk wants predictive profiles to target a solicitation based on a rental list of names, for which only demographic information is available, and the ownership of a cellular phone is not known. The offer is a discounted rate plan, which should be attractive to cellular phone owners with high monthly usage (around 500 minutes of use per month).

Even though Cell-Talk does not have the time or money to conduct another test mailing to obtain their target group of high-monthly-usage respondents, they can still develop profiles to help in their targeting efforts, as long as they have a notion of what their target group looks like. Cell-Talk can use a look-alike group — individuals who look like individuals in the target group — as a substitute for target group. This substitution allows Cell-Talk to develop look-alike profiles: profiles that identify individuals (in this case, persons on the rental list) who are most likely to look like individuals in the target group.

The construction of the look-alike group is important. The greater the similarity between the look-alike group and target group, the greater the reliability of the resultant profiles. Accordingly, the definition of the look-alike group should be as precise as possible to ensure the look-alikes are good substitutes for the target individuals. The definition of the look-alike group can include as many variables needed to describe pertinent characteristics of the target group. Note, the definition always involves at least one variable that is not available in the solicitation file (in this case, the rental list). If all the variables are available, then there is no need for look-alike profiles.

Cell-Talk believes the target group looks like their current upscale cellular phone subscribers. Because cellular conversation is not inexpensive, Cell-Talk assumes that heavy users must have high income to afford the cost of cellular use. Accordingly, Cell-Talk defines the look-alike group as individuals with a cellular phone (OWN_CELL = yes) and INCOME greater than \$175,000. Look-alike profiles are based on the following assumption: individuals who look like individuals in a target group have levels of responsiveness

similar to the group. Thus, the look-alike individuals serve as surrogates, or would-be respondents.

It is necessary to keep in mind that individuals identified by look-alike profiles are expected probabilistically to look like the target group, but not expected necessarily to respond. In practice, it has been shown that the look-alike assumption is tenable, as solicitations based on look-alike profiles produce significant response rates.

Look-alike profiling via tree analysis identifies variables that discriminate between look-alike individuals, and non-look-alike individuals (the balance of the population without the look-alike individuals). Effectively, the discriminating variables produce varied look-alike rates.

The original sample data in Table 1 are used along with INCOME to create the LOOK-ALIKE variable required for the Tree analysis. LOOK-ALIKE equals 1 if an individual has OWN_CELL = yes and INCOME greater than \$175,000; otherwise LOOK-ALIKE equals 0. There are 300 look-alikes and 700 non-look-alikes, resulting in a sample look-alike rate of 30 per cent.⁶ These figures are reflected in the top box of the look-alike tree in Figure 7. (Note, the sample look-alike rate and the original sample response rate are equal; this is purely coincidental.)

Targeting the top segment yields a cumulative lift of 333 (with a 30 per cent depth of population). This means the predictive look-alike profile is expected to identify 3.33 times more individuals who look like the target group than expected from random selection (see Gains Chart in Table 5).

A closer look at the look-alike tree raises a question. INCOME is used in defining both the LOOK-ALIKE variable and the profiles. Does this indicate that the tree is ill-defined? No. For this

Figure 6: Full tree with GENDER, OWN_CELL, CHILDREN and INCOME

Table 4: Gains chart for tree defined by gender, own_cell, children, and income
 *(Segments are ranked by response rates)

Segment*	Size of group	Number of responses	Cumulative responses	Response rate %	Cumulative rate %	Cumulative lift
#1 — INCOME, [13000,25000) CHILDREN, yes	15	15	15	100.0	100.0	333
#3 — GENDER, male INCOME, [25000,157000) CHILDREN, yes	200	100	115	50.0	53.5	178
#2 — GENDER, female INCOME, [25000,157000) CHILDREN, yes	30	15	130	50.0	53.1	177
#4 — INCOME [157000,250000] CHILDREN, yes	300	120	250	40.0	45.9	153
#6 — OWN_CELL, yes GENDER, male CHILDREN, no	200	30	280	15.0	37.6	125
#7 — OWN_CELL, no GENDER, male CHILDREN, no	200	20	300	10.0	31.7	106
#5 — GENDER, female CHILDREN, no	55	0	300	0.0	30.0	100
	1000	300		30.0		

particular example, INCOME is a wanted variable. Without INCOME, this tree could not guarantee that the identified males with children have high incomes, a requirement for being a look-alike.

LOOK-ALIKE TREE CHARACTERISTICS

It is instructive to discuss a noticeable characteristic of look-alike trees. In general, upper segment rates in a look-alike tree are quite large and often reach 100 per cent. Similarly, lower segment rates are quite small, and often fall to 0 per cent. These patterns are observed in Table 5. There is one segment with a 100 per cent look-alike rate, and three with a 0 per cent look-rates. The implications are that:

- it is easier to identify an individual who looks like someone with predefined characteristics (for example,

gender and children) than someone who behaves in a particular manner (for example, responds to a solicitation)

- the resultant look-alike rates are biased estimates of target response rates, to the extent the defined look-alike group differs from the target group. Care should be exercised in defining the look-alike group, because it is easy to include individuals inadvertently unwanted
- the success of a solicitation based on look-alike profiles, in terms of the actual responses obtained, depends on the disparity of the defined look-alike group and target group, and the tenability of the look-alike assumption.

CONCLUSION

The distinction was made between descriptive and predictive profiles. The predictive profile is used for finding the

Figure 7: Look-alike tree

Table 5: Gains chart for look-alike tree defined by gender, and children, income
 *(Segments are ranked by Look-alike rates)

Segment*	Size of group	Number of responses	Cumulative responses	Look-alike rate %	Cumulative rate %	Cumulative lift
#4 — INCOME, [176000,250000] CHILDREN, yes GENDER, male	300	300	300	100.0	100.0	333
#2 — CHILDREN, no GENDER, male	400	0	300	0.0	42.9	143
#1 — GENDER, female	100	0	300	0.0	37.5	125
#3 — INCOME, [59000,175000] CHILDREN, yes, GENDER, male	200	0	300	0.0	30.0	100
	1000	300		30.0		

firm's best customers, after which the descriptive profile is used to communicate effectively with those customers. Tree analysis was introduced, and it was shown how trees can be used as a user-friendly, powerful method of

developing a set of complex and interesting predictive profiles. The tree-based approach of profiling was expanded to include look-alike profiling, a reliable method when response information is not available.

References

- 1 In fact, a descriptive respondent profile may also describe a typical nonrespondent. We actually have this situation in Tables 1 and 2.
- 2 More likely to respond than a random selection of individuals.
- 3 This is CHAID.
- 4 The term 'large' is subjective. So is the Tree display, however, which is an inherent weakness.
- 5 Extreme segment response rates (close to 0 per cent and 100 per cent) reflect another inherent weakness of Trees.
- 6 The sample look-alike rate sometimes needs to be adjusted to equal the incidence of the target group in the population. This incidence is rarely known and must be estimated.