Market modelling · 9
Data mining models - 1

Chuck Chakrapani

The database revolution
Recent advances in computers and database marketing have created a renewed interest in target modelling. Target modelling refers to a group of techniques that are akin to market segmentation. Both approaches aim to identify customers who have more of certain characteristics (such as propensity to buy our product). However, there are major differences between the two.

Traditional market segmentation mostly works on a conceptual basis. For instance, it can identify your market segments as 'wealthy professionals' or 'socially mobile women'. But there is a limitation in this approach: it is not easy to locate 'wealthy professionals' or 'socially mobile women'. While such segmentation is definitely superior to indiscriminate mass marketing, the limitations remain.

A well-maintained database, on the other hand, precisely identifies target customers and provides the means of reaching them. It provides a means of exploiting data on current, lapsed and potential customers. While traditional segmentation techniques helped in market strategy, target modelling provides an effective means of implementing the strategy. Databases enable marketers to be customer oriented. While traditional segmentation enabled product centred strategies, database modelling enables customer-centred strategies. This shift in focus from products to the customer also coincides with the service quality and customer satisfaction orientation of recent years.

Prediction and planning
Marketers want to influence the buying behaviour of consumers in a way that is advantageous to the firm. To do this, the marketer needs to know what consumers will do. Traditional market research procedures aim to achieve this goal by measuring consumer attitudes, what will make them switch brands, what demographics are associated with higher purchase probability and so on. While such data served -and has been serving-marketers well, there are other aspects to consider. First, because surveys are expensive, the data are collected from a sample. Even a large survey seldom exceeds a sample of 2,000 consumers. This means that we do not have information on our customer, but have information that can be extrapolated to our customers. Second, our information is essentially aggregate. Third, we do not have any background information on the consumer except what we ask in the survey. Fourth, we have absolutely no information on those who do not respond to our survey. Fifth, we have no way of tracking the behaviour of a customer over a period of time except through specially setup panels, which share some of the weaknesses of surveys. Sixth, regular surveys provide only a limited means of testing the effectiveness of a marketing campaign. A good database overcomes some of these limitations and provides the marketer a means of extending the power of traditional marketing research. Many databases provide the means of incorporating sample surveys, making them richer in terms of their marketing value.

Data Mining Techniques
The data mining techniques that are used to model databases allow the marketer to go on 'fishing expeditions'. They allow the marketer to specify his or her hypothesis in fairly weak terms. For instance, a marketer can look at all the variables in the database in the hope that some of them might be related to purchase behaviour. The data mining models can uncover any hidden patterns in the data, the strength and the hierarchy of relationships. This is not radically different from traditional analysis, but the availability of suitable data and computer power takes this from what once was an onerous task to an effortless exercise.

Tree Modelling
Probably the most popular of all data mining techniques is 'tree' modelling. As an example, consider a bank with a large database which sends out a mailer every month promoting one of its products. Further assume that of the 12 mailing pieces sent out each year, only about 3 may appeal to any given customer. If this is so, then the bank

is spending 300% more than is necessary to generate a given level of response. Moreover, when customers receive frequent and mostly irrelevant mailers, they ignore them, thereby reducing the probability of responding even to a relevant mailer. It would make sense, then, to mail each piece only to those segments of customers with a high probability of responding. Instead of sending 1,000,000 pieces and receiving a response of 10,000, it might be better to mail only 300,000 pieces and receive a response of 8,000. Mailing costs drop by 70% while the response drops only by 20%.

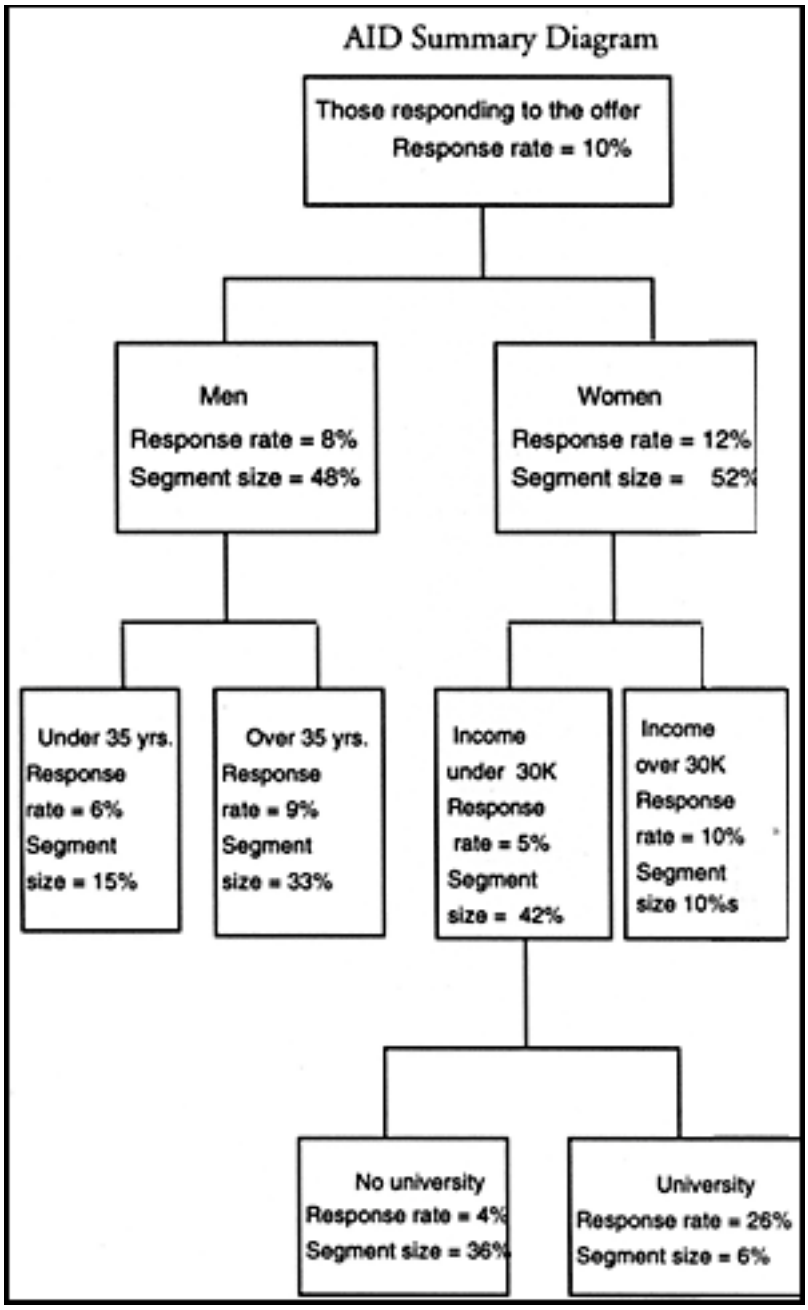
To identify high response customers for different types of products, the marketer can use past data. What are the characteristics of customers who responded to different types of mailings? The assumption is that past behaviour is predictive of future behaviour. For instance, if past data show that the response rate for an RRSP mailer is 37% among those who are 50 years or older, 28% among those who are between 30 to 40 years of age and only 5% among those who are 29 years and younger, we would expect such a pattern to hold for this year as well. Even if it does not, we are unlikely to find a dramatic shift in response rates. This is particularly so if we can find a logical relationship between the dependent and independent variables. In this instance, it is logical to expect that older people as a group tend to have more money to invest and that as people get older they show greater interest in retirement products.

But things are seldom that simple. How about the interaction between different variables? Is it having the discretionary income or is it becoming older that is the basic driver in the acquisition of a retirement product? How do marital status and children affect buying behaviour? Does income have any influence? How about educational level? If all these factors influence response levels, how are they related to one another? Which attributes are more important?

'Tree' models are designed to answer questions like these. The approach first divides consumers on the basis of the most influential variable. Then it searches the remaining variables to identify the next most influential variable. These models continue to identify variables, in descending order of importance, until some present criterion is satisfied.

An overview of Tree models

Tree modelling started with the work of Morgan and Sonquist in the early 1960s at the University of Michigan. Their technique, known as the Automatic Interaction Detector or AID, used one way analysis of variance techniques to identify the most influential variable and divide the sample into two groups. The technique is designed to split the groups two ways (binary splits) at each stage. Their model, exciting as it was, did not receive widespread acceptance in marketing research. There were several reasons for this. First, binary splits are highly restrictive. In our example, AID would not identify the three age groups with different response levels automatically. It would first divide the group into two. This could be, for example, those who are above 30 and those who are below 30. Those who are above 30 will then be split into those who are under 40 and those who are above. Which could be all right except that the second level split of those over 30 might be superseded by another variable, for example, income. When this happens the relationship between age and the response variable becomes less clear. While technically AID gave the best possible split at each level, the logic of the splits was not always clear. The second limitation was if the variable is continuous (such as income), the modeller has to convert it into a categorical variable, by grouping the actual income into different income brackets. (Newer techniques also require this, but they can do it automatically, without the modeller having to specify the groups, unless s/he chooses to do so.) The third limitation was AID required large sample sizes. AID is still being used in marketing research, but newer techniques (such as CHAID and KnowledgeSEEKER) are rapidly replacing its use. We will discuss these newer techniques in the next issue of Imprints.


AID Summary Diagram
Those responding to the offer
Response rate = 10%
Men
Response rate = 8%
Segment size = 48%
Women
Response rate = 12%
Segment size = 52%
Under 35 yrs.
Response
rate = 6%
Segment
size = 15%
Over 35 yrs.
Response
rate = 9%
Segment
size = 33%
Income
under 30K
Response
rate = 5%
Segment
size = 42%
Income
over 30K
Response
rate = 10%
Segment
size 10%s
No university
Response rate = 4%
Segment size = 36%
University
Response rate = 26%
Segment size = 6%

Book Review

Applied Multivariate Techniques

by Subhash Sharma
Published by John Wiley & Sons, New York. $85.95

Multivariate techniques ranging from factor analysis to conjoint analysis are used fairly routinely in marketing research. Wide availability of computers and prepackaged programs have brought these techniques within the reach of most researchers, even if they are not strongly quantitative. Yet most books that explain what these techniques are and how they work are strongly rooted in formulas and mathematics. This is frustrating to intelligent researchers and marketers who wish to understand the techniques conceptually and apply them appropriately. Many books with the term 'applied' in their titles tend to be less rigorous in their exposition and use fewer formulas but are sufficiently complex to be beyond the reach of an intelligent but non-mathematical reader. In effect, they tend to be inferior technical books.

This situation has not changed with this book. However, unlike many other 'applied' books, this book is written carefully. For readers who are not afraid of symbols, Sharma provides a very meaningful explanation of techniques. His examples are many and relevant. His explanation of computer outputs is clear and well thought-out. Mathematical details of the techniques are given in the appendix, so the interested reader can understand the technical basis of the techniques discussed.

Applied Multivariate Techniques covers principal components analysis, factor analysis, confirmatory factor analysis, cluster analysis, two-group and multi-group discriminant analysis, logistic regression, MANOVA, canonical correlation and covariance structure models. While the omission of some techniques like conjoint analysis can be justified on the grounds that a book does not have to cover every technique, the author's decision to omit regression analysis is unfortunate. The author's justification is that this technique should be studied separately. While the author is free to choose what he wants to write about, not including regression analysis in an applied text diminishes the usefulness of the text, since it is one of the most widely used techniques. It is hard to agree with the author's contention that it (along with ANOVA) needs a separate book. Most books cover regression analysis in a single chapter quite satisfactorily.

Notwithstanding the truncated coverage of the techniques, the reviewer feels that this is a good book. This book is not for the casual reader, but is useful for someone who is not mathematically sophisticated but is not afraid of formulas and equations . Sharma's explanation of computer outputs is one of the best I have seen in any multivariate analysis book. This is an applied text written with care, a relatively rare phenomenon these days. Recommended for all serious users of multivarite analysis.

Chuck Chakrapani

Dr. Chuck Chakrapani of Standard Research Systems is a Toronto-based consultant, author and seminar leader. He works internationally.