Using External Data Sources to Optimize Data Mining Solutions

by Richard Boire

Historically, the purchase of external data sources has always occurred at a list level. The purchase of data at this level (i.e. as a list source) has been the traditional and simplest way of enhancing direct response.

For instance, the use of the *Financial Post* list would optimize the response of a mutual fund product while a *Sports Illustrated* list would achieve the same effect for a sporting product. Obviously, we can increase responsive behaviour by buying more targeted but smaller lists.

However, the trade-off of less names will always present a challenge when we are trying to both maximize orders as well as minimize costs. Besides the revenue limitations of a targeted list, the other limitation is the inability to discriminate between high responsive and low responsive names within a targeted list. List sources discriminate at a group or list level rather than at an individual level.

The incorporation of statistical tools and in particular, modelling, has allowed the industry to overcome both of these above limitations. Modelling can be applied to a huge list of names as well as provide the capability to discriminate individual names by their response behaviour.

The advantages of this kind of technology have enhanced the evolution of database marketing and its impact within the direct marketing industry. As the complexity of database marketing increases, the need for both more and better quality data is the key to building better segmentation tools for direct marketing programs. Because of this need, new companies are being formed to specifically address this demand. In order to differentiate between these kinds of companies, a technique for properly evaluating data sources is required.

This evaluation will assess how our direct marketing segmentation results are improved. These segmentation results are manifested in two areas:

1. Increased orders
2. Reduced mail costs

The benefit of any data source will be adding incremental value in either or both of the above-mentioned areas.

The ability to increase orders is achieved through better profiling which can be obtained by a number of statistical techniques such as cluster analysis, etc. Better profiling is attained through the identification of different groups of people and different communication strategies to each group based on the group behavioural needs.

The ability to reduce mailing costs is also achieved through a number of statistical techniques such as regression modelling. Its direct effects are that we can obtain the same number of orders at a lower mailing cost thereby resulting in a lower cost per order.

Let's take a hypothetical example of a data source called XYZ and how we would evaluate it.

Before we even assess the XYZ source, we must examine our own in-house database. For the sake of simplicity, we are a financial services firm which markets its insurance products through direct marketing. Its database information consists of the following:

1. Name and address
2. Postal code
3. Number of purchases
4. Tenure with company
5. Age

It is important to understand our existing information since it would be meaningless to undertake any type of analysis if an outside data source contained no new information.

Furthermore, we should also determine how each piece of data is collected (i.e. at the postal code or individual level) since individual-level data provides better *segmentation results*. For XYZ, the information looks as follows:

1. Name and address
2. Postal code
3. Income at postal walk level
4. Number of credit card purchases
5. Own or rent home
6. Age

A preliminary assessment of this data source will reveal that age will be useless since it already exists within our database while income is irrelevant because it is available from Statistics Canada at the postal walk level. Therefore, the relevant pieces of information from this data source are number of credit card purchases and own or rent home.

Assuming our annual direct marketing program for this financial services company consists of a million names, an average response rate of three percent, a premium price per policy order of $50 and a mailing cost of $0.90 per name, we can now calculate the value of this data source in terms of increased orders and reduced mailing costs.

The first approach of assessing value based on increased orders would utilize cluster analysis or some other statistical technique. One group of cluster groups would be created with the new data while another group of clusters would be created without the data. The offer or message is then designed for each cluster. A test mailing is then sent out to both groups of clusters. The results of this test mailing reveal that the response rate for clusters with the data is 3.5 percent while the response rate for clusters without the data is three percent which is the overall response rate of the entire mailed population. The value of this data for increasing orders is calculated as follows:

(3.5%-3.0%) x 1,000,000 x $50 = $250M

incremental response due to data source	total mailed names	price per order	value due to increased orders

The second approach of assessing value based on reduced mailing costs would essentially determine the incremental gain provided by this data to our existing scoring models. For instance, if 1,000,000 names represents 50% of our total available universe, then we need to determine the impact of modelling with and without this data at 50% of our available universe.

	models without data source	models with data source
50% of universe	3%	3.3%

The value of reduced mailing costs is calculated as follows:

(1,000,000 – (.03 x 1,000,000)/.033) x $.90 = $82M

present mailing names without data source	required mailing quantity needed to achieve same number of orders with data source	mail cost per name	value of reduced mailing costs

The total value of this data source is $250M + 82M = $332M. If the cost of the data source

is $100M, the purchase of this would be a high priority.

It is evident in this example that increasing orders is the more significant contributor to the data's value. However, this is not standard for all situations. In some cases, we may find that reduced mailing costs is the more significant contributor.

This approach provides a consistent way of measuring the value of data from different vendors. As companies strive towards improved cost efficiencies, the ability to utilize the best available data represents a clear example of a continuous process for improvement.

Richard Boire is a principal partner at the Boire Filler Group, a data mining consulting company.