Deconstructing Data Mining: The key to customer-level profitability

By Richard Boire

Data mining has become one of the most misunderstood terms in the direct marketing world. Although the definitions are numerous and sometimes conflicting, they all agree that data mining represents the detection and identification of previously unknown customer trends. Data mining does indeed produce these results, but it also involves much more. It represents the process of intelligently using customer-level data for profit maximization within future direct marketing campaigns. Although this sounds like an overly simplistic definition, the data mining process contains a progression of specific detailed stages which are designed to achieve the overriding goal of campaign profitability. I have identified four major steps or stages which are required in any data mining project:

1. Problem Identification;
2. Creation of the analytical file;
3. Application of the appropriate technology; and
4. Implementation and tracking

Problem Identification

This is the most critical stage of any data mining project. Having the best statisticians and best technology in the world is irrelevant if the problem is not identified correctly. This may seem like an easy stage, but in many cases it is the most difficult one. Experienced data mining practitioners refer to the communication gap between marketers and data analysts as the major obstacle in identifying the problem. This communication gap is quite common for most organizations today. Industry pressure is creating the need to train and educate people who can bridge this gap. A new growth industry is being created as companies strive to hire personnel who have both a superior understanding of direct marketing and its impact within an overall business strategy as well as a superior understanding of the data environment and how it can be used to optimize direct marketing programs. This component of both technical and business knowledge is the key towards successfully identifying the business solution.

Creation of the Analytical File

This step in many cases is the most labour-intensive. With the problem identified, data can be pulled from many different files and systems, and integrated into one overall file. This requires an understanding of the many data elements which are required given the established objective. At the same time, programming skills are usually employed in order to combine all this information into a single file. Before actually combining all the information into one file, a data audit is performed on all potential data elements which indicates the usefulness of the information. For instance, a data element containing a large proportion of missing values is irrelevant. In other cases, the values of particular fields may be combined into categories to form more meaningful variables. Only once the data audit process is complete can all the information be integrated into one analytical file.

Application of the Appropriate Technology

The nature of the business problem and its resolution dictates what kind of technology should be used. In many cases, the technology model can be as simple as producing a list of postal codes ranked by customer penetration. In more complex situations, such as building customer lifetime value for credit card companies, numerous models are required. The tremendous growth in data mining software has facilitated the modelling process by allowing analysts to employ a variety of statistical techniques. Techniques such as correlation analysis can be used to statistically prioritize the importance of variables against a given predicted behaviour. In the actual development of the final model, techniques such as CHAID, regression analysis, neural net models, etc. can all be employed. All statistical techniques are equally valid in building a predictive model. The real determination of the best technique for a model is not dependent on the statistics but how the model performs when applied to a validation sample. Sorting a validation sample by descending model score, we can then bucket the validation sample into 10% increments ranging from 0-10% (highest score names) to 90-100% (lowest score names). The objective of the model is to then optimize rank ordering based on actual performance. The issue of what statistical technique to use can then be based on whether the technique can significantly improve the actual response rates within the top two deciles.

Besides identifying what the appropriate modelling technique is, analysts as well as marketers should have a solid understanding of the components or variables within the model. In other words, modelling should not be a mysterious black box solution. It should provide a basic understanding of why certain variables are or are not in the model. Exploratory data analysis reports are created to provide this kind of analytical capability. These reports demonstrate the impact of a certain variable against the predicted behaviour. Listed below are two examples of analysis with one report leading to a useful variable and another where the variable was of no use:

MODEL VARIABLE

Age	Response Rate
Under 25	5%
26-35	1%
36-45	1.5%
46-55	2%
55+	2.5%

NOT A MODEL VARIABLE

Years as a Customer	Response Rate
Under 1 Year	.75%
1-2 Years	1.25%
3-4 Years	1.00%
4-6 Years	0.90%
7+ Years	1.05%

From the above reports we can discern that older persons are more likely to respond which would indicate that this variable would be positive within the overall equation. Meanwhile, no discernible trend emerges within tenure. As a result, we would not expect this variable to be in the model.

This example provides just one small sample of the analysis which can be produced using technology. The key in using technology for data mining solutions should be the enhancement of analytical capabilities.

Implementation and Tracking

In many cases a data mining project fails due to this last stage. The first three steps may be executed with great success, yet a project will always fail due to improper application of the learning. For instance, a classic case of misapplication occurs when the overall database changes between the time of learning and its implementation. Suppose that age was the most important model variable determined during the time of development, but perhaps privacy restrictions caused it to be removed completely from the database. Any model application would probably result in sub-par results since this piece of information could not be used in the generation of a targetted customer list. If the analyst recognized this, the best advice would be to delay the campaign until a new model is developed.

Besides implementation, a tracking system should be designed to determine if this learning is improving overall campaign performance. This learning should be integrated into the overall campaign objectives. Specific tactics are then adopted to implement this learning with results which are easily monitored through a sound tracking system.

This last stage of the process is critical in determining whether the existing data mining solutions still perform or whether new solutions need to be developed. As a result of this stage, we see that data mining is not a one-shot process, but an ongoing process used to provide solutions for constantly evolving businesses.