

CANADIAN
MARKETING
ASSOCIATION

CMA



ACM

ASSOCIATION
CANADIENNE
du MARKETING

Best Practices in Data Mining

Executive Summary

Prepared by: Database & Marketing Technology Council

Authors: Richard Boire, Paul Tyndall, Greg Carriere, Rob Champion

Released: August 2003

Executive Summary

Canadian marketers have long used data mining as an important tool to help improve the effectiveness of their marketing campaigns, and over this period some organizations have emerged as industry thought leaders in data mining.

To enable more Canadian marketers to benefit from the effective use of data mining, the Canadian Marketing Association's (CMA) Database & Marketing Technology Council established a sub-committee to conduct research and report its findings to the CMA membership. The subcommittee then considered the research approach (structured interviews plus gathering of anecdotal evidence), established the questions to be employed, and identified target organizations. The interview subject companies were selected from those in: Retail; Financial Institutions; Publishing; Not-for-Profit; Telecommunications; and Technology, as well as Data Mining Specialists.

The research phase was conducted during the summer and fall of 2002, and consisted of 15 formal interviews, most conducted in person, and lasting more than one hour. Interview subjects were granted anonymity to encourage candor and thereby gather the widest possible range of experiences. Additionally, the interview questions were structured to reflect the process of a typical data mining project.

Perhaps the survey's first significant finding was that the process of defining and understanding the business problem at the core of every data mining project is not a straightforward, linear process. Instead, we found that gathering information and developing a suitable analytical approach defies a neat, sequential description. Instead, information requirements and analytical development overlap each other and typically require multiple iterations.

In addition, we found that the longer an organization had been involved in data mining, the larger and broader the stakeholder group having influence in defining the problem and gathering the information. In fact, there is a strong correlation between the length of a firm's data mining experience, and the size of the group involved in information gathering and exchange. In our survey, those indicating many sources of information outnumbered those with a more ad hoc approach (and quicker passage of this phase of the project) two-to-one. As a firm's experience with data mining grows, so does the extent of its information gathering. However, there does not appear to be a direct correlation between the size of the organization and the extent of its formal processes.

Prioritization of data mining projects is a significant issue for Canadian marketers, as resources are not infinite. Almost half of respondents indicated some formalized return-on-investment calculation or process is used to determine priorities, although there is significant variation in the use of formal metrics versus forecasts, estimates and assumptions. Firms with longer experience were more formal in their use of return-on-investment calculations. Another large group of respondents (38 per cent) indicated that priorities are set in the context of meetings held to discuss the organization's needs, with resolution of remaining issues escalated to a higher management level.

Although well-organized, accessible data is the foundation of effective data mining, our respondents noted that accessing and manipulating corporate data may require as much as 80 per cent of a data analyst's working time, while analysis itself occupies just 20 per cent. We observed that the extent and complexity of managing data effectively is frequently a function of the organization's growth and success. A financial institution, for example, could easily have multiple data files for a single customer, each recording the customer's information a slightly different way – a product-centric approach. A customer-centric approach is preferred, in which one customer record is home to all product or service relationships.

Most interview subject companies were currently between the extremes of product-centric and customer-centric, and financial services companies had made most progress integrating their data. Financial institutions also appear to have done the best job documenting data to make it more usable.

Hardware and software considerations indicated a strong preference for industry-standard database systems, while some firms made use of desktop PC-oriented databases, including Microsoft Excel and Microsoft Access.

SAS and SQL were the two most frequently reported data extraction tools. While extraction format appears to be a matter of the end user's preference, data extraction is customized to the needs of a given project.

Many respondents also reported continuing interest in assuring data integrity (through audits and other means) as well as determining appropriate means to link data files intelligently to construct analytical files.

Clearly, the most advanced firms in data mining have placed some consistent emphasis on ensuring that data is accessible, well organized and of high quality, while some have even automated, to some extent, the consolidation of customer data.

In conducting an analysis, our interview subjects stressed the need for data mining personnel to have excellent analytical and interpretive skills to be able to extract meaning, relevance and importance from mathematically expressed information. Our subjects have not gone so far as to recommend that data miners have advanced degrees in mathematics or statistics, but that they should have excellent business sense on a strong analytical foundation.

In this regard we found universal acceptance of SAS analytical tools, with support for SPSS and their Enterprise Miner and Clementine suites, respectively. Best-practice organizations also frequently supplement their use of SAS tools with automation assistance and with more specialized statistical analysis tools. More leading-edge organizations also get right down to the programming-language level in their use of FORTRAN and C++.

Data mapping and illustration tools (such as Mapinfo) also help marketers understand geographic relationships, such as between store locations and customers' homes, or to illustrate the degree to which a given product has penetrated customers' homes in specified neighborhoods.

Data miners also use the standardized reporting tools provided in software applications such as Cognos. The output from these applications can then be loaded into Excel, using pivot tables to provide the ability to drill more deeply into the data. Such tools have practical use in analyzing consumer spending across, for example, income, gender and region.

Best-practices organizations are led by personnel with the analytical capability required to assess solutions, combining both art and science. Mathematical tools employed by these leaders include: cluster analysis; factor analysis; correlation analysis; neural nets; decision trees; linear and logistic regression; and newer techniques such as genetic algorithms and fuzzy logic.

Many of the best-practices organizations with which we spoke are producing their own algorithms to determine customer-level profitability and, thereby, customer-level ROI. Tactical, product-specific models enable organizations to target the right customers for the right product that ultimately increase that customer's value to the organization. We encountered some organizations that have developed more than 100 such product or channel-specific algorithms. Web-based modeling appears to lag behind the rest of the industry as it has generally been unable to link data with specific customers.

The final phase in a data mining project is implementation. Here, data audit is viewed as a critical component to ensure seamless data transfer between parties and to identify changes in the environment during development of the solution and its application. Our interview subjects view measurement and tracking capability to be essential in being able to communicate findings throughout the organization.

Our subjects further point out the increased likelihood of 'something going wrong,' which reinforces the benefits of automation during this phase. Consequently, data miners are increasingly turning to new campaign management products and suites.

In addition to responses received in answer to questions on data mining projects, we also solicited our survey participants' input on other subjects, including: Research & Development; Overlay Data Sources; Statistical Techniques; Software Tools; Organizational Structure; and, the Technical Environment.

The following points are presented in summary:

- There is no correlation between a firm's size and the extent of its information gathering;
- Almost half of respondents (46 per cent) indicated some form of Return on Investment calculation;
- The best practice organizations all operate in the PC environment;
- Virtually all best-practice organizations use SAS as one of their primary analytical tools; many also supplement SAS with special-purpose tools;
- Data mining groups typically report to VP or SVP management;
- Best-practice companies have larger groups performing data mining, averaging 10 people directly involved;
- Two best practice organizations have developed (in-house) their own tools to automate less analytical tasks;
- The time required to perform various forms of analysis does not vary significantly across either experience or industry; and,
- Even the most advanced data mining practitioners conduct very little evaluation of alternative techniques.