CANADIAN
MARKETING
ASSOCIATION
CMA

ACM
ASSOCIATION
CANADIENNE
du MARKETING

Best Practices in Data Mining

White Paper

Prepared by: Database & Marketing Technology Council

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

Released: August 2003

Table of Contents

Objective

Data mining has long been employed by marketing organizations to improve their marketing efficiency. Over that time, some companies have emerged as industry thought-leaders. And, even though data mining has been available for many years, some companies have just begun to explore its potential in their organizations.

This paper seeks to provide a guide for companies – both advanced and those just starting down the data mining road – to allow them to understand how advanced data mining practitioners are differentiated from those less advanced. Interestingly, a number of commonalities across data miners of all skills and experience provide encouragement for those just starting out. Readers should be able to take away tangible and actionable lessons to help improve their data mining efforts.

Methodology

Our first step was to create a subcommittee from the Canadian Marketing Association's Database & Marketing Technology Council to spearhead the initiative. The group's initial discussions determined the appropriate methodology. After determining that the research portion of the project should take the form of a series of in-person interviews, the logical next step was to develop the interview questions. We broke these questions down into categories mirroring the data mining process itself:

- Identification of the Business Problem/Challenge
- Creating The Analytical File
- Conducting the Analysis
- Implementation

We also included a section called 'Other Issues' to capture additional insights that did not fit neatly into this process. One key objective of the questionnaire design was to ensure that it would result in specific insights and findings that could lead to actionable steps for readers.

The next step involved identifying companies to be included in the interview process. We decided that we needed to cover a range of industries, company sizes and experience levels to capture the greatest number of potential differences. This would also help to ensure that the end result was relevant to all CMA members. Industries included in the study included:

- Retail
- Financial Institutions
- Publishing
- Not-for-Profit
- Telecommunications
- Technology

- Data Mining Specialists

We then targeted companies in each industry for interviews, and this resulted in a large number of candidate companies. Subcommittee members eventually performed 15 interviews over the summer and fall of 2002. Each interview lasted between 1 and 1.5 hours and was generally conducted in person by a subcommittee member. To ensure that the research generated the largest number of insights, we have allowed the interviewees and their respective companies to remain anonymous.

The subcommittee then compiled the results into a standard template and used that as the means to generate this final report.

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.

Identifying the Business Problem

Introduction

The first step in any Data Mining project is the identification of the business problem, which data mining will be used to solve. Without careful attention to the business issue and objectives at hand, all downstream analysis may be misdirected. At best, this could result in sub-optimal recommendations or, at worst, wasted effort and the reduction of business value.

For the purpose of this survey we have broken down the identification of the Business Problem into three stages:

1. Information-Gathering
2. Objective Setting
3. Prioritization

While these are critical components of identifying the business problem, they are not necessarily sequential; in fact there is often a great deal of overlap and many iterations across these stages.

Information-Gathering

Succinct identification of a business problem requires input from many sources. The business issue needs to be solved in the context of the organization’s technological, process and capabilities structure. For an appropriate assessment of the proposed path to addressing the problem, some form of cost-benefit analysis is required, whether conducted formally or otherwise. As many stakeholders are typically involved in the success of any data mining project, it is key, first, to seek out detailed information across the affected areas (typically including, but not limited to, Marketing, IT, Channels, Product Mangers, Suppliers and Business Executives).

Organizational, technical, and executional dimensions are reviewed in the process of evaluating potential impact versus effort required. Information gathering and analytical approach development are conducted in an iterative fashion. As the information is reviewed, an approach begins to crystallize. As the approach further crystallizes, the next iteration of informational needs become clear.

In our study, 62 per cent of respondents indicated there were many sources and stakeholders involved in the information-gathering stage for setting the data mining approach. A further 31 per cent indicated a more ad hoc approach that sometimes involved a few sources, with a comparatively quick passage through the information-gathering phase.

We found a very strong correlation between the length of time a firm has been involved in data mining and the depth and breadth of stakeholders’ involvement. It is clear that as organizations' experience with data mining grows, so does the extent of the information gathering.

On the other hand, there appears to be no correlation between the size of a firm and the extent of information gathering. Of the firms we surveyed, there are both larger and smaller firms found in each of the broad and narrow categories of information sourcing and exchange.

Objective Setting

While information gathering sets the boundaries and framework for the approach, it also allows for the objective of the analysis to be clearly set. Our survey had two primary interests in how data mining business objectives were set:

- Was the data itself used to help set objectives?
- Who had primary responsibility for setting the objectives?

First, with respect to using the data to help set objectives, we noted several key findings. We found that almost half of the respondents (46 per cent) used data in some way, or at some points, to assist with objective setting. Of these, none said it was a consistent practice. On the other hand, most respondents mentioned that it was their goal to use data more frequently in the objective setting process.

As before, there was a strong correlation between the length of time a firm has been using data mining and its use of data in setting objectives. Firms that had been using Data Mining longer were more likely to use their data to assist in the objective setting process.

Secondly, with respect to responsibility for setting objectives, there were a few interesting results. A substantial majority of respondents indicated that the data mining teams were part of the objective setting process (85 per cent). That is, the team had some degree of input and influence in deciding the needs and metrics of the data mining projects (versus taking direction from another business department).

There were varying degrees of involvement for the 85 per cent of respondents who were part of the objective setting process. Slightly less than half of this group, 45 per cent, indicated a primary role in determining objectives. The remainder were more likely to be sharing or alternating responsibilities with an external business area.

In both the objective setting and information gathering areas of the survey, Data Miners in the Financial Services sector are consistently more sophisticated and more involved in the Identification of the business problem.

Prioritization

Given that organizations have limited resources with which to conduct data mining, a key question is deciding where to apply resources. We found substantial variation in the firms' processes for prioritizing projects.

A few organizations used a rigorous planning and analysis approach to setting priorities. Priority setting started with an annual planning process, driven from the organization's stated objectives. From there, data is used to help identify key opportunities expected to have significant impact on enterprise goals.

These key opportunities are then assessed for the benefits to be achieved compared to the resources available. At this point additional resources can be identified and allocated (or not) on the basis of a business case. At this stage the planning process typically involves executives across several business units.

This process then sets the data-mining group's high-level plan for the year. As the year progresses the plan may be revisited often for updates on potential impact, resources or new initiatives to help re-align future efforts.

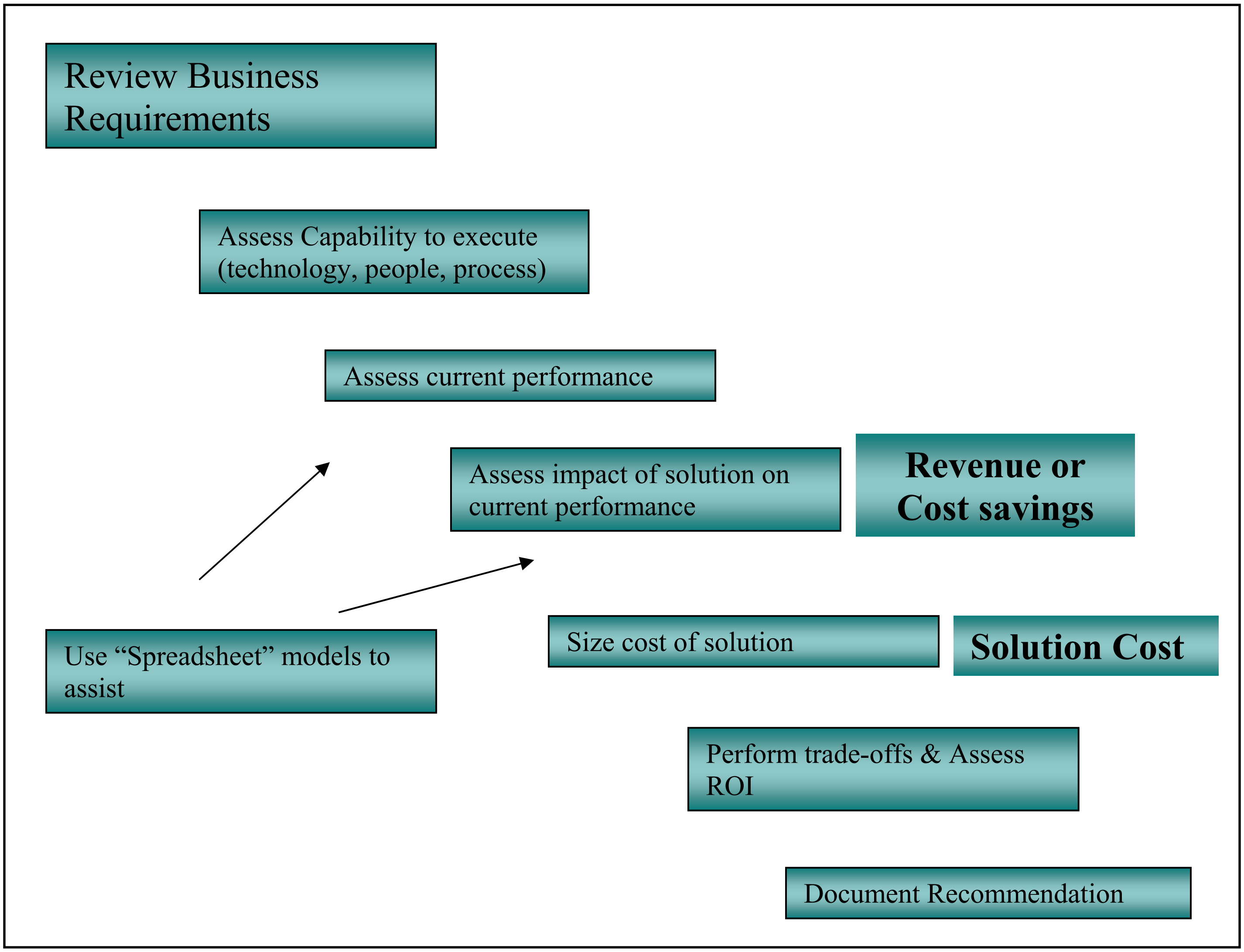
Several of the organizations use Return On Investment (ROI) calculations to assist in prioritization. In some cases, historical data is used in a very sophisticated and intensive analysis to forecast project ROI. In other cases, ROI is projected, but based more on informed estimates and assumptions. In either scenario, priorities are established using some combination of data and science.

In our survey, 46 per cent of respondents indicated some use of ROI calculations to assist in prioritization. Once again, this was more prevalent in firms that had been performing data mining longer.

In 38 per cent of respondents, the most common process for prioritization was holding a meeting to discuss perceived organizational needs. Judgment is used with the business users community and the data miners to address potential short- and long-term impacts, and some projects are selected to move forward. Such meetings were fairly regular across several business areas, and prioritization decisions not resolved are escalated for decision at a higher management level.

Most of the respondents were working to improve their prioritization processes, and there was general agreement that priority setting needs improvement. In some instances, prioritization was determined by which business unit was screaming loudest.

A high-level illustration of a prioritization process is shown below:



Review Business Requirements

Assess Capability to execute (technology, people, process)

Assess current performance

Assess impact of solution on current performance

Revenue or Cost savings

Use "Spreadsheet" models to assist

Size cost of solution

Solution Cost

Perform trade-offs & Assess ROI

Document Recommendation

Creation of the Analytical File

Introduction

It may be a generalization but, in many organizations, accessing and managing corporate data takes 80 per cent of the analyst's time, while truly analytical work takes only 20 per cent. Accessing and managing data can be a daunting task.

Well-organized accessible data is the foundation for advanced analytics, such as data mining.

The Challenge of Data Management

In a complex environment, data resides in multiple sources, across multiple platforms, and often with different variable definitions and data standards. To bring data together from these disparate sources requires an in-depth understanding; it also requires the knowledge that the data actually exists (i.e. you have to know what data is actually available), and an understanding of its limitations. Another requirement is time – organizing data can be an extremely lengthy task.

How does this relate to our study of data mining and this section devoted to the creation of the analytical file? What are the best practices in this area, and how do they allow an organization to confront some of the challenges described above?

An example to illustrate the point:

First, recognize how disparate data sources evolve. As it grows, an organization maintains more and more information on its customers. Different departments may have developed their own databases, and different product lines may have developed standalone databases as well. Due to these different sources of information, a single customer may be seen as several distinct customers:

- Joe Smith – credit card customer
- JP Smith – mortgage customer
- Joseph Smith – student loan
- J Smith – savings account…and so on

Second, imagine the challenges faced by the person who must access this data, match it and integrate it. In the example above we are taking a product-centric view and changing it a customer-centric view. Here is how we should now see this customer:

- Joseph P. Smith
  - Credit Card Account
  - Mortgage Account
  - Student Loan
  - Savings Account

Now imagine that these data sources are on different platforms, in different parts of the organization, and managed by different groups. Your analyst may be denied access to the information, but he or she may request Database Administrators (DBAs) to access copies of this data, and this could take time. As likely, there may also be multiple *real* J Smiths in the customer base, adding to the complexity.

These are some of the basic steps required <u>before</u> analysis can be started.

Now add to this the overlaying of external data.

Add to this the issues associated with dirty data, multiply by the size of your customer base, multiply again by the number of data sources you are accessing, and you have a formula for a very time-consuming task.

While many financial institutions may have moved forward with data warehouses and data marts that reduce or eliminate this problem, it is a widespread phenomenon across business types; wherever there are multiple customer data sources, there may be a significant challenge in turning this data into a file ready to be used for data mining (or even more basic analysis).

<u>What's the ideal?</u>

In the ideal scenario, an analyst would have access to all customer data. It would be cleaned, de-duplicated, and overlaid with any desired external sources. From here, the analyst could move quickly forwards with data mining. The analyst would essentially have the ability to take an extract of data, and work with it directly.

To make this possible, a datamart or data warehouse would be used where data from all the disparate data sources are already tied together and refreshed on a regular basis.

<u>The reality</u>

The organizations we spoke with find themselves somewhere in between our fictitious example of Joe Smith's bank and the ideal scenario above. Data issues will never be eliminated – but they can be largely overcome, leaving the analysts available to do their best work, rather than spending valuable time gluing together various data.

<u>Research Findings</u>

As part of the survey we asked a series of 20 questions on this segment of the data mining process.

Data accessibility:

In almost all cases, access to customer data was not an issue. In one case, there were internal data sources identified, which were not being made available, presumably as a result of privacy considerations or competitive restrictions. The only other instance of inaccessible data was that of a not-for-profit organization, which hosted its data externally with a third party. (This third party performs data mining on the customer database).

The best practicing organizations have the data in a warehouse or datamart. This is the starting point when there are multiple sources to consolidate. This should provide one source for customer data, after which it may be overlaid with external data sources.

Documentation of data:

To create and manage a datamart or data warehouse, the content and structure of the data must be well documented – the term for this documentation is metadata – again, the best practicing organizations have this information and make it available to the analysts.

Hardware and Software:

In terms of specific hardware and software environment, generally industry-standard Relational Database Management Systems (RDBMSs) were mentioned. These included Oracle, SQL, NCR Teradata and others. Other organizations mentioned use of desktop tools such MS Excel (spreadsheet) and MS Access (database).

The key issues with databases are to have well organized data and a place to keep it. The brand of database would depend on other needs, such as size of file, performance requirements, security, and similar considerations beyond the scope of this survey.

Data Extraction:

The two tools used most commonly to extract data were SAS and SQL. SAS can access most databases on most platforms. SQL is a well-known tool for querying and extracting data.

Extraction format appears to be a matter of preference. Commonly mentioned were ASCII, text and comma separated value (CSV).

Survey respondents often cited a standard list of exclusions. Exclusions would include, for example, customers who have defaulted in the past and who should therefore not be offered further credit. Another way to view exclusions is as those individuals whom, for some reason you do not want to contact: they may already be customers of the product or service you are promoting.

Data Content and Enhancement:

In some cases the entire file is used for analysis, but in most instances a standard subset is used, as dictated by the size of the whole file and the types of analysis tools, as well as by hardware considerations and the statistical approach. There may not always be advantages to using the whole file.

The length of history pulled varied from 13 months to more than seven years, with most respondents indicating an extraction pull of two to three years of history. The length of history available also varied. In some cases a full history was available, in others data would be archived after a certain time.

Data audits were conducted by all best practicing organizations in any project that required the use of a new source of data. The data audit essentially consists of a frequency distribution on each field for the data that is being analyzed. Listed below is an example of a data audit on the gender field within a customer master file.

Gender	# of records	% of records
Male	35000	42.17%
Female	28000	33.73%
Missing	20000	24.10%
Total	83000	100.00%

The above type of report enables the analyst to obtain a deeper understanding of the data environment. At the same time, it can be used to highlight data integrity or quality issues.

In creating new variables, one also needs to understand the data structure to link all the disparate files into one analytical file properly. This requires that the user understand the key link fields between the files. Furthermore, the user needs to understand the exact relationship between the two files. For instance, do the two files link on a one-to-one basis, a one-to-many basis, or a many-to-many basis. This understanding is critical to summarizing and aggregating information into broader derived variables. A good example of this is the ability to create a total 12-month spend variable at the customer level through a summary of 12 months of transactions. Another good example of a derived variable is the merging of very detailed sales-level information into broader and more meaningful sales groups. Data summarized at a broader sales group level will, in most cases, have much more relevance within any data mining exercise than the more detailed sales level information.

Conclusion

From the results of this portion of the survey, we can see that best practicing organizations have their data well organized, and easily accessible. By automating the consolidation of customer data (and other external data sources), analysts can focus more of their efforts towards the analysis of customer behavior.

Conducting the Analysis

Introduction

Within the other sections of data mining, we observed that the key to success is having the right human resource skill sets. Yet the growth of data mining has resulted in tremendous evolution of new tools and technologies that now comprise the data miner's toolkit. This is especially significant when conducting an analysis to build the appropriate solution. The best practicing organizations are more fully able to leverage this toolkit by having personnel with extensive knowledge and experience within this area. Yet what does this mean? If knowledge and experience are the keys, then how do they translate into finding the skill sets that work best within this area?

One essential key to success is the ability to deal with numbers and measures. More importantly, the person needs to make business sense of the numbers as an actionable solution by the business. The individual does not need to have an advanced degree in mathematics or statistics to be successful, but a strong mathematical aptitude or quantitative-type background provides a distinct competitive advantage.

In understanding the best practices within this area, the first consideration is whether the analysis and subsequent solution is statistically or non-statistically based. The best practicing-organizations within this area obviously have the full capabilities to conduct both these types of analyses at any level of detail. For example, a simple non-statistical analysis might be the ranking of postal codes by customer penetration. A more complex non-statistical analysis might be the creation of cohort-type analytical reports whereby one can view the year of acquisition for a given group of customers and observe their resulting behaviour over a number of years.

Year	# of New Donors Acquired	# of total Donors In Year	% Active	# of Gifts	Avg # of Gifts per New 1998 Donor
1998	1,054	1,054	100%	1,141	1.08
1999		316	30%	498	1.58
2000		301	29%	625	2.08
2001		272	26%	755	2.78
2002		148	14%	424	2.86
Total		2,091		3,443	3.27

An example of a simple statistical solution might be CHAID – that is, a decision-tree type solution with the end nodes or branches of the tree representing customer segments. These customer segments are determined through a statistical routine (CHI-Square) whereby segments are created based on how statistically significant they are from the desired mean customer behaviour. The customer segments from this output could represent groups included in or excluded from a given marketing campaign. An example of a complex statistical solution might be the creation of models using an advanced mathematical routine such as neural nets and/or genetic algorithms.

Key Criteria

Once the specific type of analysis is determined, we need to identify if the organization has the necessary infrastructure to support a given analysis. For instance, to address this issue the basic questions are:

1. Do we have the hardware?
2. Do we have the software?
3. Do we have the expertise to perform the required analysis?
4. What kind of reporting and communication capabilities are available?

1. Do we have the hardware?

This could be as simple as determining whether or not we have enough computers to perform certain analyses. However, at a more granular level, we also need to understand whether or not we have the required computing power. This means that we need to understand the processing and RAM capabilities of a given machine. It is typical to discover that more advanced mathematical algorithms really test CPU capacity. In evaluating new technology and software, a major consideration will always be the performance of a given analytical routine.

The best practice organizations all operate in the PC environment and would certainly have to be cognizant of the space and processing requirements. Sometimes these organizations will perform functions within the mainframe environment. It would appear that Oracle is a common platform amongst these organizations for mainframe use.

2. Do we have the software?

The most important technical decision within the entire data mining process is the selection of the appropriate software. In our research, we have discovered that virtually all of those best-practicing organizations have SAS as one of their primary analytical tools. The leading competitor to SAS in this area is SPSS and, indeed, one best practice organization has selected SPSS as its primary analytical tool. In particular, the Clementine suite of modules (SPSS) would compete directly with the Enterprise Miner suite of modules offered by SAS. Both Enterprise Miner and Clementine attempt to eliminate (or at least minimize) the programming required to develop a given solution. This is especially relevant when the analyst is creating the analytical file. With both Clementine and Enterprise Miner the entire data mining process is modularized with front-end graphical interfaces guiding the analyst through this process. Although programming intervention is minimized, the analyst must understand the process of how he or she will build the solution. Equally important, the analyst needs to create the data environment to support the development of the solution. This requires that the analyst have a deep understanding of which fields and files to use, as well as how to link these files in creating the analytical file.

We also find that best practice organizations often supplement SAS with other tools, either because SAS lacks the functionality or, as in most cases, a given functionality can be provided more cost-effectively. For example, it is not uncommon to see SAS used as the analysts' programming choice for organizing, summarizing, and aggregating data into the analytical file. At that point, the best practicing organizations will certainly use the statistics module within SAS to develop a given solution. However, these organizations will also import the SAS dataset into other statistical software to test different algorithms to improve the overall solution. Some of the more common examples of different software used to supplement SAS or SPSS for a given solution include:

- Answer Tree within SPSS
- Knowledge Studio within Angoss

More leading-edge organizations will use specific programming languages such as Fortran and C++ for some analysis. One direct result of this competency is the ability to conduct conjoint analysis to determine statistically the optimum number of test cells for a marketing campaign.

Some tools within the best practicing organizations' toolkit relate to data visualization and the ability to create reports. For instance, Desktop Mapping tools such as Mapinfo provide valuable information such as the distance between where a customer lives and the organization's location, and how this can impact marketing behaviour. Mapinfo can create maps that depict the customer penetration around a given company's location. In many cases, the customer points around a given location can be color-coded based on some demographic such as age.

Tools such as Cognos are often used to generate the standard data mining reports required by all organizations. Using SQL-type syntax, the tool essentially creates pre-built queries (cubes) based on marketing needs. These needs should be flexible so the marketer can create his or her own custom reports beyond those offered as standard. However, the ability to generate these custom reports is itself dependant on the information contained within the pre-built cube or query. If the required information goes beyond what is available within the cubes, then the marketer needs to generate a request to an SQL programmer in order to generate a new pre-built query or cube.

Other analytical software is now available to allow marketers to create these custom reports even if the information is not contained within the pre-built queries or cubes. Using database technology that relates to the optimal use of indexing in the processing of data, these products can provide the following benefits:

- Broader access to information by non-technical users
- Increased speed in creating the analytical file
- Enhanced capabilities in deriving new information.

Once the variable dimensions (or required variable classification levels) are created through some advanced data mining software, tools such as Excel and Access provide the capacity to present results in a spreadsheet or database format. The use of pivot tables within Excel provides the capability of drilling down on the information by being able to analyze certain behavio~~u~~rs across a variety of dimensions.

For instance, if we wanted to analyze spending across dimensions such as range of income, gender, and region, the user, through the use of pivot tables, could easily generate summary reports across different values within each dimension.

As data mining continues to evolve, software development will continue to provide new products for the data miner's toolkit. Much of this development will occur in new algorithms in statistics or mathematics meant to support high-end power users. Those best practicing organizations will certainly evaluate this high-end software, but will also evaluate other software which is not mathematically intensive but which does provide benefits in other areas. From the best practitioner's standpoint, the major benefits in evaluating new software are:

- Increased Targeting Capability
- Increased Speed in Conducting the Analysis
- Broader Access to Information by more End-Users

3. Do we have the expertise to perform the required analysis?

Best practice organizations have expert competencies in both mathematics and general business. The key to these organizations' success is the ability to leverage both skill sets in developing the solution. It is this integration that allows organizations to develop a data mining culture where solutions encompass both art and science. Within such organizations, the data-mining department is led by a person with a strong technical bent but, more importantly, even stronger business acumen. In these departments, the people working on a solution will have strong technical and mathematical skills. Yet the real key to success is how this work is managed or directed. This means that the data mining leader needs to understand the business problem clearly to determine the appropriate approach and tools in the development of the solution. More importantly, the data-mining leader needs to understand the results and metrics from a given solution that will determine success. With these insights, the leader can then provide the proper direction to the technical and mathematical personnel who are actually hands-on in building the solution.

Mathematical tools used by analysts in these leading organizations include:

- **Cluster analysis**
  This is often used in cases commonly referred to as unsupervised learning. In unsupervised learning, segmentation or targeting is based on the characteristics that are most common within each cluster. There is

no marketing behaviour to be optimized.

- **Factor Analysis**
  This is also another example of unsupervised learning, which provides data reduction capabilities by reducing the number of variables into a more meaningful set of fields for the analysis.

- **Correlation Analysis**
  This is often used as a preliminary statistical tool to identify variables that may be relevant for a given analysis.

- **Neural Nets**
  These techniques are used in both unsupervised cases to build clusters as well as supervised cases to build predictive models.

- **Decision-Trees**
  Techniques such as CHAID and CART are often used to produce key segments that can be either selected or excluded for a particular marketing campaign. Other uses of this technique consist of the insights being able to create new variables based on the interactions between variables within the tree.

- **Linear/Logistic Regression**
  The most common and more traditional techniques used to build models are linear regression and logistic regression. Linear Regression is often used as the technique to optimize a given variable that is continuous, such as spending or amount of credit loss. Logistic regression is used for scenarios whereby the desired behaviour is yes/no. Examples of this include defection and response rate models.

- **Other techniques are:**
  - ❖ Genetic Algorithms
  - ❖ Fuzzy logic

Many of leading-edge organizations are producing algorithms to generate customer-level profitability and, as a result, customer-level ROI. At the same time, tactical product-specific models enable organizations to target the right customers for the right product that ultimately increase the customer's value to the organization. It is not unusual for these leading-edge organizations to have developed and applied more than 100 statistical models and algorithms. In many cases, models are both product-specific as well as channel-specific. However, the experience in building these types of solutions even for the leading edge organization is minimal (in many cases, less than five years) since the industry for the most part is in its infancy.

Beyond model-building capability, these organizations display great expertise and valuable insights in being able to derive new variables and information that would create additional value in the development of any solution. Examples of this type of creativity include the development of trend- or pattern-related information as well as customer-related indexes based on key purchase or demographic metrics.

The emerging trend in data mining is the development of solutions on the web, commonly referred to as web mining. Our interviews have indicated that currently, no organization in the Canadian marketplace is really doing any customer-related Web mining. Some organizations have used tools like Web Trends to determine click-through rates concerning page views and time of day. Such tools can also be used to measure purchase rates, as long as there is an on-line purchase activity. In many cases these tools lack the ability to relate this information to individual customers and, hence, the ability to target different customers based on their Web click-stream behaviour. One organization, though, has built e-mail models. Their experience is using click stream data to provide additional lift within these models proved unsuccessful.

4. What kind of reporting/communication capabilities are available?

As part of the analytical process, many organizations employ standard CRM-type reports to obtain knowledge on the general health of their business as it pertains to customers. Using Key Business Indicator (KBI) metrics, these organizations will design reports to allow the organization to observe changes in its customer base and in customer behaviour.

One critical component in developing solutions is the ability to increase the understanding within the organization through reporting. We observed that the development of data audit reports was a critical component in the creation of an analytical file. In developing solutions, Exploratory Data Analysis (EDA) reports help to depict the impact of certain information for a given solution. The best examples of this are the EDA reports that are used in the development of a model. Shown below is an example:

Behaviour Score	**Response Rate**	**% of All Customers**	**Response Index**
800+	2.00%	33%	200
600-800	1.00%	33%	100
under 600	0.50%	33%	50
Average	1.00%	100%	100

The above EDA represents a response model variable. In the model, we found that behaviour score has a positive impact on response. The EDA helps us understand this relationship visually as we observe that there is a positive trend between behaviour score and response rate. These types of reports can be very useful in opening up the so-called ‘black box’ of modeling by providing a greater understanding about the model’s actual components or variables.

Other reports relate to how we measure the performance of a given solution. This must be done in a manner whereby the impact can be measured in business terms such as ROI. For instance, the report below commonly referred to as a Gains Chart depicts the level of performance from a model that was applied to a validation sample.

% of (Ranked Model)	Validation Mail Quantity	Cum. Response Rate	Cum. % of all Responders	ROI
0-20%	4,000	2.00%	40%	50%
20-40%	8,000	1.60%	64%	20%
40-60%	12,000	1.40%	84%	5%
60-80%	16,000	1.20%	96%	-9%
80-100%	20,000	1.00%	100%	-3%

As you can see from the above, the Response Rate model was developed off one sample and then validated against another sample. This validation sample was scored with the model and then ranked into quintiles in the above example - (normally deciles) by descending model score. The response rate results represent the observed response rates within each decile. The prime determinant of model performance is how well the model rank orders based on observed response rate. From this chart, we can see that the top 20% performs twice as well as the average. Assuming that we have promotion costs and the profit margin of the product being purchased, we can then translate these numbers into ROI (Return on investment). The real utility of this type of report is the establishment of a decision-making process based on a business impact such as ROI.

Implementation

Representing the last phase of data mining, this phase can pose large challenges and, at times, be somewhat daunting as organizations attempt to take action on the solutions that have been built. It is very common to observe data mining projects where the approach and tasks in the previous phases have been successfully executed. In effect, these result in a very sound data mining solution. But the implementation of the project neglects a certain detail that causes the solution to be totally ineffective.

Because of the required detail, a rigorous data audit approach is also the recommended course to ensure that solutions are being implemented. The data audit approach in this phase attempts to accomplish two main objectives:

- Transfer of data between parties is seamless, in that loading of files and fields is similar for both organizations
- Identify significant changes between what has happened during the development of the solution and its current application

Another key component in successfully implementing solutions is the ability to create a proper measurement and tracking environment for the solution. Although we may have successfully applied the solution within the current data environment, the lack of any measurement and tracking system mitigates against being able to communicate results to key stakeholders within the organization. This can be very significant since it is these stakeholders who are the key influencers in any future funding or investment decisions for business initiatives. Inability to communicate results could certainly diminish future funding within the data mining area.

The increased likelihood of something going wrong further accentuates the need for automation within this phase. In fact, we will see that the best practicing organizations attempt to maximize the level of automation, including such processes as:

- Loading of data
- Scoring/List Creation
- Tracking/Measurement

The attempt to automate the tracking and measurement component has created a whole new suite of products and services around the notion of campaign management. Such services (in theory) allow all end users (technical and non-technical alike) to design and execute the creation of complicated marketing matrices. Specifically, end-users could generate specific lists or groups for the various test and control cells within the matrix. At the same time, they could create specific campaign codes that would be a source of information in the generation of any standard campaign reports. Such reports would then be produced within a specific period subsequent to the campaign.

Standard promotion history reports allow the end user to track the impact of promotions for specific campaigns and over a certain period of time. The information derived from these reports allows the user to define overall corporate strategies around the frequency with which the customer should be contacted.

The analyst could also generate ad hoc reports that might be specific to a campaign rather than the standard knowledge that would be required after every campaign. For instance, the tracking of a model represents an element of information that is only going to be useful if a model is applied during a campaign. Listed below is a schematic of what this might look like.

Model Deciles (ranked by model score	# of Names Promoted	Response Rate
1	50000	4.00%
2	50001	3.00%
3	50002	2.50%
4	50003	2.25%
5	50004	2.00%
6	50005	1.50%
7	50006	1.25%
8	50007	1.00%
9	50008	0.75%
10	50009	0.40%

This type of information can be extremely useful, particularly when a campaign has been deemed unsuccessful. Suppose in the above example, that we needed an overall response rate of 5 % to break even in this campaign. Obviously, this campaign does not generate the acceptable response rate. Initial thoughts on understanding why the campaign did not generate 5 % might simply be that the model is not working. However, by producing the above decile-tracking report, we can indeed demonstrate that the model is an extremely effective targeting tool by its ability to rank order response rates from the highest decile to the lowest decile.

In assessing the best practices of organizations, we looked at five tasks:

1. Technical Implementation
2. Creation of Files/Data
3. Tracking/Testing of Solution
4. Distribution of Solution
5. Post Campaign Analysis

1. Technical Implementation

The ability of an organization to implement solutions internally is dependant on both the number as well as the complexity of tools to be applied. Essentially, if the requirement to extract information for a solution is weekly, then it makes sense for an organization to internalize this process. The internalization of this process would require that organizations have resources to conduct the following tasks:

1. Apply data mining/scoring instructions to database to generate appropriate list extract for campaign.
2. Generate reports to validate relevancy of solution to current environment (e.g., determining how model score ranges have changed between time of development and implementation).

2. Creation of Files/Data

This varies by client. However, standardized file formats are the norm for best practicing organizations producing this information internally. For data and information going outside, standardized routines will be in place to produce acceptable file formats.

Ultimately the best-practicing organizations seek out solutions to facilitate the creation of files and data. This means an increased degree of automation to provide enhanced capability to more business users. For instance, these tools should allow a non-programmer to create a list and to do it very quickly – often with same-day turnaround – if the request is not too complex. Yet these tools should have sufficient flexibility to enable end-users to generate lists for a wide variety of business scenarios.

We found that the most common software for generating a file applied with a data mining solution is SAS. Other tools have been used and are as effective, yet the history of SAS as the primary tool for experienced end users makes it the preferred tool.

3. Testing/Tracking

One of the key issues emerging from this area is the level of involvement from the analytical group in designing and creating the appropriate test and tracking measures. Here, the most successful organizations have strong communication channels between the analytical group and the business end-user group. A formal process regarding this communication dialogue is established from the outset with an initial meeting to obtain consensus on the common goals and objectives to be achieved from the campaign. This consensus is critical, as the business group and analytical groups bring very different perspectives to the meeting. The business user will have keen insight on his or her business expertise, and how it affects the bottom line. However, the analyst will have a high degree of expertise in whether or not the data exists to achieve end user’s desired campaign objectives. More importantly, the end user’s knowledge of the data environment is critical as to how the testing and tracking matrix will be created.

Another best practice is having the ability to translate this information into return on investment. It is understood that a number of different performance metrics will be evaluated during the course of a given campaign and in fact they may differ depending on the type of campaign. Yet, in all cases, there should be a process to translate these metrics into ROI.

4. Distribution of Solution

The types of channels that organizations used during the course of their campaigns were:

- Direct Mail
- Telemarketing
- Sales Force
- E-Mail/Internet

We found that the use of data mining was quite minimal in the E-mail - Internet environment. We believe the virtual zero cost of E-Mail - Internet marketing has been a mitigating factor regarding the use of data mining in this medium. Yet we fully expect this to change as it matures, and becomes a more developed channel for all organizations. In particular, the best practicing organizations are now employing some Web mining solutions but, more importantly, are attempting to determine how to use the information gathered from Web mining to target more effectively in the more expensive channels.

It also appears that the best practicing organizations are shifting more of their channel costs from direct mail to telemarketing. These organizations use both channels and are constantly assessing the optimal proportion of these channel costs within their overall budgets.

5. Post-Campaign Analysis

Virtually all organizations use their analytical teams to produce and distribute the results. However, the best-practicing organizations have the skills and resources to produce both simple and complex backend analyses. For instance, a simple backend analysis may be the reporting of tracking results from cell codes within a matrix. A more complex backend analysis may involve a comprehensive review of a model's performance within a campaign. This may involve two steps:

1. Assessing the Model Variables
2. Assessing the model performance and its impact within the various communication options

<u>Post-Campaign Analysis</u>

Shown below is an example of a report that assesses the model variables within a post-campaign analysis.

Model Variable	Impact on Response during model development	Impact on Response during current campaign	Stat. Sign. At 95% within current campaign
Tenure	negative	negative	yes
Previous buyer	positive	positive	yes
Live in Toronto	positive	no impact	no
Income	positive	negative	no
# of persons in household	negative	positive	no

From this report, one can surmise that there are problems with the current model for three of the five variables. Two of the variables exhibit a different impact between time of development and the time of the current campaign. A third variable no longer has any statistical significance within the current campaign.

Shown below is another example of a report depicting a model's actual performance within the campaign versus expected performance.

Model Decile	Response rate performance during model development	Response rate performance during the campaign
1	8.00%	5.00%
2	7.00%	4.25%
3	6.00%	3.75%
4	5.00%	3.25%
5	4.00%	2.75%
6	3.50%	2.00%
7	3.00%	1.65%
8	2.50%	1.25%
9	2.00%	1.00%
10	1.50%	0.80%

In this example, we see that the model is still effectively rank-ordering response from the top decile to the bottom decile as it did during the model development. However, due to the much lower overall response rate, we can surmise that the campaign itself may be the problem and not the model.

The above illustrations demonstrate just some of the more advanced types of tracking that would be utilized by organizations that deploy models within their current campaigns.

Besides producing the reports for tracking, the analytical group will also offer insights and key findings concerning these reports. For instance, in the case of a model where three of the five variables are no longer relevant, further investigation by the analytics team would be required to discover why this has happened. This type of exploratory analysis further reinforces the notion that the analytics team will play more of an advisory role within the best practicing organizations rather than being mere messengers of information.

In their continual efforts to improve the effectiveness of accessing information, many leading-edge organizations have purchased software that attempts to eliminate much of the manual effort in generating reports. This means that much of the programming that may have been required in the past to generate reports is now automated. Once again, this enhanced level of automation empowers the broader business community to generate its own reports, without the intervention of any programming support.

Other Considerations

In addition to the procedural elements reviewed in this research, there were also a number of items with a broader, but no less tangible impact on the data mining process. In this section we will discuss our findings with respect to these other considerations, as they relate to the data mining carried out by our target companies.

Research and Development

While most business people typically associate research and development with less practical applications than data mining in the marketing environment, we found that there is still a role for it. Data mining research can fall into many categories, including research into new data sources, new statistical techniques and new software tools.

The key finding from this area of the interviews was that while most agreed that R&D could lead to increased efficiencies, companies spend very little or no time in this area. All companies interviewed, whether advanced or beginner, spend the majority of their time executing data mining analysis to support current strategic and tactical objectives.

Data Sources:

Data miners are always looking for ways to enhance their customer profiles to increase the power of their analysis. In addition to leveraging their internal data sources, they often look to external, third party data sources to extend their understanding of their customers and prospects. In Canada, we are severely limited in the number and variety of data sources available to append to customer files. Data miners working on Business-to-Consumer (B2C) or Business-to-Business (B2B) applications are faced with the same limitations. This is due partly to Canadian legal policies limiting the resale of consumer data to third parties, and partly due to the size of the Canadian market for such products. In the U.S., less restrictive laws and the larger marketplace mean that substantially more detailed information is available on consumers, often at an individual level.

Databases such as Experian and Acxiom provide detailed financial assessments of their prospective consumers for rent by marketers. In Canada, this data is much more limited and usually provided only at an aggregate level, meaning that it can only be used to profile the average characteristics of the neighbourhood in which the target prospect or customer lives.

Examples of some of the most commonly referenced overlay and prospecting data sources include:

Consumer Data

- National Census – published every five years, available at Enumeration Area level
- TaxFiler – compiled from annual income tax returns, available at postal walk level
- ICOM – twice yearly consumer survey

- Neighbourhood Cluster systems e.g. Mosaic, PSYTE
- Loyalty program data – e.g. Air Miles, Aeroplan
- Subscription lists – e.g. newspapers, magazines
- Neighbourhood level credit scores

Business Data

- Dun & Bradstreet
- CBI
- InfoCanada
- Statistics Canada
- Subscription Lists – industry publications

Our findings from the interviews suggest that the use of overlay data sources is not as dependent on the analytical sophistication of the companies as it is on the required applications. Some beginner and intermediate companies used overlay data as much or more than their more advanced counterparts. This is somewhat offset by many advanced practitioners having much more extensive internal customer databases and not needing to rely as heavily on external data. In our findings, however, even the most advanced companies usually evaluate both external data and internal data to improve the effectiveness of their data mining efforts.

Statistical Techniques

Data mining applications can employ a variety of statistical techniques. Many of the current software tools provide lengthy lists of techniques to be tested and evaluated. However, in our findings, companies did not spend significant time pursuing these alternatives. Instead, the vast majority of data mining analysis consisted of the core statistical techniques of regression, CHAID and clustering, in addition to non-statistical techniques such as triggers, decision rules and profiling. And within regression analysis, the most commonly used technique was Logistic, which is specifically designed to address Yes - No situations such as: did a customer respond or not? Even the most advanced practitioners did very little evaluation of alternative techniques.

One of the advanced practitioners regularly benchmarked its predictive models against university research methods. This provides the assurance that it is producing the best analysis on a continuing basis. However, it should be noted that is not a common approach and that only one company uses this method.

Software Tools

This is where companies focussed most of their R&D efforts. Most companies mentioned that they try to evaluate new software tools as they become available. However, no company, advanced or otherwise, has a formal procedure for tracking new software developments, or testing their efficacy in their environment. These are mostly limited to ad hoc Beta tests and product evaluations. The core toolsets are generally those employed most often for day-to-day projects. This reflects our finding that most practitioners use a small set of core analytical techniques to accomplish most of their tasks. The current tool sets on the market are generally sufficient to deliver these capabilities in an efficient manner.

Organizational Structure

The structure of the data-mining group can have a significant impact on its ability to produce analysis effectively. Its overall position in a company's organization will affect the likelihood that a company will effectively leverage the potential for data mining to drive marketing effectiveness and, with it customer profitability. For data mining to be truly effective in an organization, it must have high-level support.

This means that senior people in the company must believe in the power of data and the ability of the data mining group to leverage this data effectively. For data mining, the results of the analysis must be implemented to make a positive impact on the bottom line. And this implementation often requires some degree of process change, such as which lists to target, how the sales force prioritizes its client contacts, or which customers should receive discounted pricing. Without sufficient support, the effective implementation of these results could thereby be jeopardized.

In all but one company surveyed, the data mining group reports to the VP or SVP level, primarily in marketing or sales. This clearly indicates that the data mining discipline is widely accepted in Canada as a valuable tool for marketers. Again, there was little difference between beginners and more advanced practitioners in this area.

Often the more advanced practitioner companies have larger groups doing data mining. This does not mean, however, that large groups are required, as the average advanced group had approximately 10 people directly involved in data mining. There were two exceptions to this, as one advanced practitioner reports having 70 staff while another reported that more than 100 people are involved. The beginner and intermediate practitioners had only a few people involved.

Another alternative is the outsourcing of the data mining function. Several of the less advanced practitioners outsourced either parts or all of their core data mining functions to external vendors that specialize in data mining and analytical services for marketing applications. None of the advanced companies reported that outsourcing was an alternative staffing solution that they pursue regularly.

Technical Environment

Automation

We found that the use of automated tools is limited. Some elements that automated data mining tools combine include: data processing and standardization, sampling, multiple statistical techniques for predictive modelling or other technique, evaluation and scoring.

Generally, these were created from the combination of distinct tool sets and sometimes offer additional functionality.

Two of the advanced practitioners have developed their own in-house tools to automate less analytical tasks including recurring model scoring and standard variable creation in the creation of the analytical file – repetitive, straightforward tasks suited to automation, as they can easily be standardized and basic tests can flag potential issues. As well, removing the responsibility for performing these mundane tasks from the analytical staff frees them up to focus on more valuable work and leads to greater job satisfaction. However, none of the interviewed companies stated that they are using these automated tool sets for performing data mining projects. Most personnel performing data mining are experienced with the tool sets, and have the ability to execute the majority of the required data processing and analytical tasks, and prefer to maintain control over the flow of the process.

Web mining is the one area where some companies focus their automation efforts. This is again likely closer to the scoring and variable creation model whereby standard routines and decision rules can highlight key findings or issues. Due to the large volume and on-going influx of data associated with use of the web, this likely explains why the focus is on automation. Most companies have yet to pursue this area to any degree, but several state that they expect this will be a growth area.

<u>Limitations</u>

During the course of the interviews we asked people what they felt were the biggest obstacles encountered while trying to achieve their data mining goals: Technology, Data, People and Training. The result was surprising.

Almost every company felt that they did not have enough personnel to complete all of the data mining projects that had been identified or they felt would add value to the company. Due to the combination of technical skills and business acumen required to be successful in data mining, this is a significant issue. Most would agree that there is a limited talent pool available from which to draw for these individuals. And, since most companies feel they need more, this should be a serious concern for companies looking to take advantage of data mining.

A much smaller portion of companies, around 20 per cent, felt limited by their technology, whether hardware or software. In general, the advanced practitioners felt they had sufficient resources available to complete their data mining tasks. It was the less advanced companies that felt technology was a bottleneck for them. This is interesting in that it could be more of a perception than a reality. Based on the question about lack of people, it is unlikely that acquiring new technology will move these companies up the continuum.

Only 15 per cent of companies felt that a lack of data, internal or external, was a substantial limitation to their data mining goals. This is in keeping with the generally held belief that data miners can generate results with whatever data is available. Experienced data miners would always like to have access to more information, but there is rarely a situation in which there is not sufficient data available to perform some type of analysis. More data will generally lead to better solutions but data miners learn to trade off between ease of access and cost of data versus the potential value to the final solution.

In general, most of the companies surveyed felt that adequate training was available to help them perform data mining effectively. One of the more advanced practitioners stated that only limited training was available, but did not really consider this to be a limitation. They believed that once data miners have gained a certain degree of experience, formal training is less effective than informal training gained through exposure to new projects and analytical techniques.

Timing

Companies were asked to provide general information about the duration of common types of analytical projects. One of the key findings here was that the timing did not vary significantly across experience or industry. There is, however, a high degree of variation in what would typically be included within certain types of projects, particularly in the less statistical exercises of customer profiling and campaign analysis.

For a data mining project such as predictive modelling, the timeline to develop the model was generally in the three-to-four week range, with little variation across companies. Profiling analysis had a much wider variation, but would typically be completed in one or two weeks, although this would depend on the number of customer groups or segments to be profiled, and the complexity and depth of the data used to derive the profiles. Campaign tracking again had a wider range due to the variation in what would typically be included in an analysis. However, the timing was generally in the same range as a profiling analysis, averaging one to two weeks. For respondents executing a large number of similar campaigns, standard reporting templates and simplified campaign metrics can easily drive this time down to hours or days. However, for more complicated campaigns done on a less frequent basis, one to two weeks is the standard.

Key Findings

The following section highlights some of the key findings uncovered during our comparisons of the most advanced practicing data mining organizations and those less advanced.

Sixty-two per cent of respondents indicated there were many data sources and stakeholders involved in the information gathering stage for setting the data mining approach.

The length of time that a firm has been involved in data mining and the depth and breadth of stakeholders' involvement are strongly correlated.

There is no correlation between the size of a firm and the extent of information gathering.

Almost half of the respondents (46 per cent) used data in some way to assist with objective setting.

Eighty-five per cent indicated that the data mining team is part of the objective setting process, with 45 per cent having a primary role in driving out objectives.

Forty-six per cent of respondents indicated some use of ROI calculations to assist in the prioritization process.

The best practicing organizations have the data in a warehouse or datamart, and it is well documented.

The best practice organizations all operate in the PC environment.

Virtually all best practicing organizations have SAS as one of their primary analytical tools

Many best practice organizations supplement SAS with other tools, because SAS either lacks a certain functionality or that functionality can be provided in a more cost-effective manner.

Some best practice organizations have developed tools in languages such as Fortran and C++ to help with data visualization and reporting.

The use of overlay data sources is not as dependent on the analytical sophistication of the companies as it is on the required applications.

Even the most advanced practitioners do very little evaluation of alternative techniques.

No respondent has a formal procedure for tracking new software developments and testing their efficacy in their environment.

The data mining group generally reports to the VP or SVP level, primarily in Marketing or Sales.

Best practice organizations have larger groups of people doing data mining with an average of approximately 10 people directly involved in data mining.

While the use of automated data mining tools is fairly limited across all companies, two best practice organizations have developed their own in house tools to automate less analytical tasks including recurring model scoring and standard variable creation in the creation of the analytical file.

Almost every company felt limited by a lack of data mining personnel.

Twenty per cent of companies felt limited to some degree by their technology, whether hardware or software.

Fifteen per cent of companies felt that a lack of data, whether internal or external, was a substantial limitation toward their data mining goals.

Most companies felt that adequate training was available.

The time to perform various types of analysis does not vary significantly across experience level or industry.

Glossary

To find definitions of many of the terms used in this paper please refer to the Database glossary contained on the CMA's Web site at the following location: http://www.the-cma.org/council/databasecouncil/glossary2002.pdf