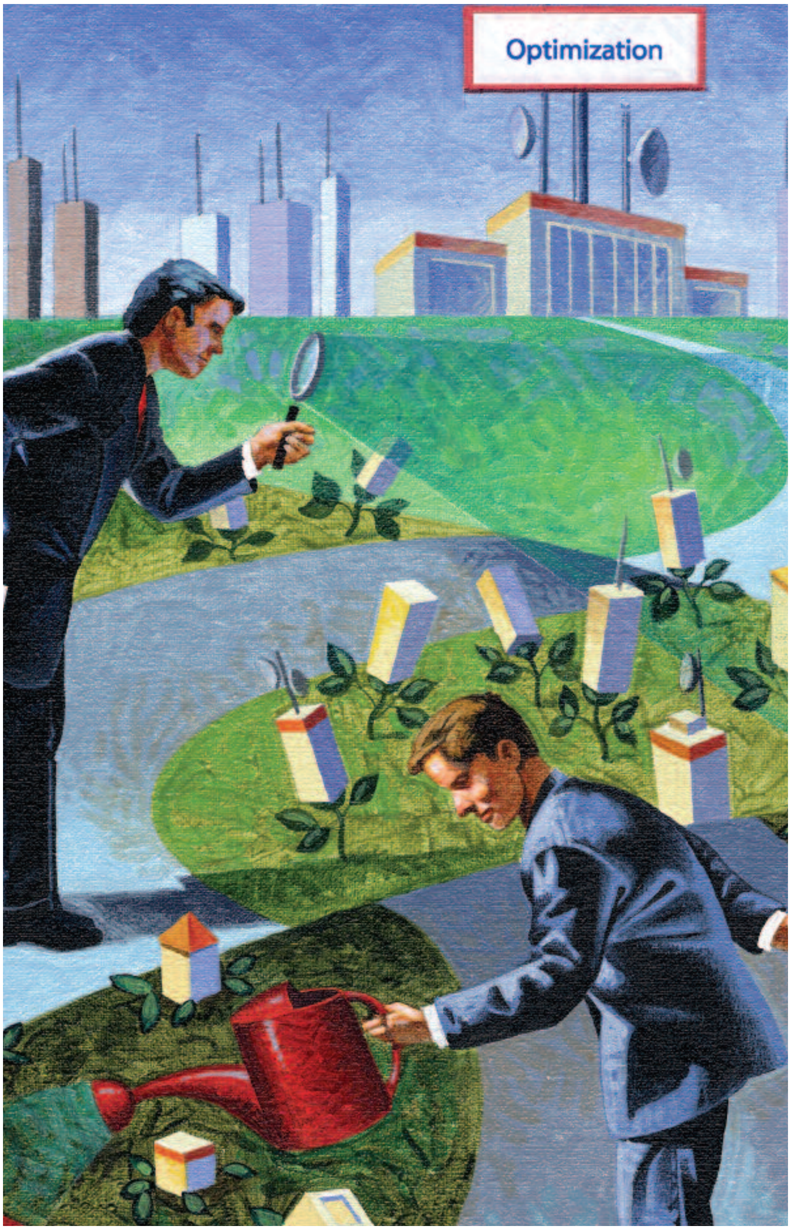
DM Review

Information Is Your Business

Volume 15, Number 10

www.dmreview.com



data mining

How to Buy Data Mining

A Framework for Avoiding Costly Project Pitfalls in Predictive Analytics

How to Buy

A Framework for Avoiding Costly Project Pitfalls in Predictive Analytics

By Eric A. King

How does someone purchase an intangible, cryptic, seemingly immeasurable technology? Beyond the inherent up-front risks of engaging in what is essentially a discovery process, just identifying a starting point can be intimidating and mystifying. Despite its elusive nature, data mining technology has surpassed the flash-in-the-pan "miracle tool" stigma with widespread and sustained success stories highlighted in mainstream publications, along with recurring case studies of improved operational efficiencies, enhanced business intelligence and residual payback. For any organization with annual revenues more than $50 million, employing data mining technology is not a matter of whether, but when.

Data mining has been seeping into mainstream business applications for more than two decades. Numerous case studies may be quickly referenced via a simple Internet or publication search. Its progress is unstoppable, propelled by sustained value justifications - yet stinted by the complexities of development, interpretation, integration and adoption. This article will suggest how to properly approach the starting line and

Data Mining

how to implement a purposefully flexible framework for establishing an efficient and effective organizational data mining process.

Cutting Through The Buzz

Let's first make sure that we're on the same page when talking about data mining. It is not wholly incorrect to label data mining as retrospective searches on a large database for specific criteria, otherwise known as online analytical processing (OLAP) or SQL queries. An example of OLAP or SQL queries would entail mining a large repository to identify females between the ages of 28 and 45 from New York, New Jersey and Delaware with incomes between $65,000 to $90,000 who purchased blue slacks between July 1 and August 15. For this query, we know the exact question to ask of the database. This practice typically explores just 5 to 15 percent of a large database.

For the purposes of this article, data mining shall refer to computer-aided pattern discovery of previously unknown interrelationships and recurrences across seemingly unrelated attributes in order to predict actions, behaviors and outcomes. Simply put, when referring to data mining in this article, we are looking at prediction derived from information hidden within large volumes of data rather than retrospection drawn from an OLAP or SQL query.

It is important to recognize and relate much of the popular terminology that is thrown about in order to provide context going forward. Data mining technology is not new. Methods for automating pattern discovery and prediction have existed for decades. Despite a considerable level of hype and strategic misuse, data mining has not only persevered but also matured and adapted for practical use in the business world. How could a community that is so data-rich, yet information-poor and profit-driven abandon a tool that can validate its own ability to predict customer behavior?

Alongside the technology, terminology has evolved over the last four decades. Names from 40 years ago are still recognizable as common phrases today. In the '70s and '80s, names such as artificial intelligence and machine learning that implied the computer had its own consciousness were somewhat oversold (perhaps even "over-souled"). The names of various data mining building blocks such as neural networks, genetic algorithms and evolutionary computing deservedly carry Darwinist tones of natural selection, as the underlying mathematics emulates biological processes. From a mathematician's perspective, these processes may be viewed as statistics on steroids.

In the '90s through the early '00s, the technology has been commonly referred to as data mining and knowledge discovery. However, due to the duality of the term data mining often referring to both OLAP and pattern discovery, a shift is rapidly moving toward far more descriptive and accurate nomenclature such as predictive modeling and predictive analytics. In fact, this will probably be one of the last articles I write using the label data mining - which is too mainstream to abandon just yet.

Just What *is* Data Mining?

Is data mining considered a service? Is it hardware? Software? A scored file? A system or a process? A customized solution? There does not seem to be a consensus, which makes data mining all the harder to visualize, define, manage ... and purchase. Two people may discuss data mining and have entirely different concepts in mind. Of course, all of the previously mentioned descriptions are technically correct. While the business community may appropriately view data mining as a productive, value-driven solution, that perspective focuses on the destination, not the journey. If credit were given to the best definition of data mining, process would score the point.

Viewing data mining as a process encompasses all the hard and soft resources, and implies a structured yet ongoing approach to an evolving optimization problem. When viewed as a process, data mining projects may be planned and implemented in a procedural way that all but ensures success. As well, expectations should be inherently leveled to never expect a "final answer" nor anticipate a single pass. When implemented properly, productive results should be expected early and continually improved.

How Not to Buy Data Mining

It is far too common for organizations to adapt their data mining project design to a blend of their perception of what data mining is with a standard corporate practice for evaluating and purchasing products and services. The result is a popular yet doomed approach:

1. Collect product literature from data mining tool vendors at industry events or as advertised in journals.
2. Invite vendors whose retail price of their flagship product fits within available discretionary budgets to visit on site.
3. Gain a free education in data mining through subjective presentations at the vendor's expense (too many are anxious to chase any sales bait, qualified or otherwise).
4. Purchase a data mining tool from the vendor who presented last.
5. Throw some data at the tool and await magical results.
6. Stare at the numbers or even visualizations thereof, wondering why an angelic chorus did not accompany the results.
7. Without knowing whether the results are useless or phenomenal, data mining is dismissed as hyped and/or pie-in-the-sky technology.

The ultimate cost of a failed first pass can be tremendous. Not only will the organization suffer opportunity costs from value never realized, but competitors will also have a greater window to capitalize on the benefits. Furthermore, morale will be adversely affected, which can wreak untold havoc on any organization.

Ultimately, data mining will be utilized by all medium and large organizations in one form or another. Not employing predictive analytics against a large repository of data (which all medium and large businesses have) would be analogous to building drilling platforms, pipelines and storage tanks with no intentions for a refinery. Although the term data mining may fade, the technology will not. If a company makes a failed approach now, it will only need to repeat the attempt later. The question is whether the organization will repeat its mistakes.

A Best-Practice Approach to Data Mining

The recommended approach for data mining presented in this article has a perfect track record of matching performance to expectations. Data mining is essentially a discovery process, which requires a purposefully flexible framework with numerous checkpoints for assessment and adjustment. Be wary of any vendor who proposes to deliver a fully implemented data mining solution without early decision points. Consulting firms with reputable names will often win sizable data mining contracts and proceed with a weak strategy free of checkpoints and adjustable stages. The project quickly migrates into an exercise of post-justifications, blame casting, contract-wiggling and backpedaling.

The following five stages provide a foundation to drive a successful data mining strategy and implementation.

1. Training

The best results in data mining are achieved when a data mining expert combines experience with an organizational domain expert. While neither needs to be fully proficient in the other's field, it is certainly beneficial to have a basic grounding across areas of focus. Even if a data mining project is entirely outsourced, substantial advantages await the organization whose principals are trained to recognize elusive pitfalls, speak confidently about data mining methods, appreciate tradeoffs between accuracy and explainability, collaborate more effectively for data preparation and interpret more accurately the model's results. Such knowledge can also serve well toward evaluating vendors, interacting with project managers and effectively questioning any suspect results or methods.

Numerous data mining conferences and public training courses exist. Many tool vendors have excellent instructors and worthwhile courses, particularly for their customers. Most times, however, courses offered by tool vendors restrict the scope of content to highlight the capabilities of their product(s). Because tools should not be considered until later in the process, try to identify vendor-neutral conferences and courses to receive an unbiased, broad and nonpromotional presentation.

If staff or time simply does not exist to train internally, consider hiring an independent data mining expert who may act as a liaison and third-party project advocate between your organization and the main project vendor. The consultant should hold three qualities in combination:

1. He or she should be well-steeped in the data mining process with a strong

track record of application success.

2. He or she should be multilingual - able to converse fluently with analysts, IT staff, users, directors and executive management.
3. Most importantly, he or she should be business-oriented, not rushing to analyze the data, but focusing first on amassing a comprehensive understanding and assessment of the client's business model and all available resources, as well as any applicable history, benchmarks and objectives.

Whether conducting your project internally or outsourcing, it pays to incorporate the direction of a data mining expert. Working in concert with an organizational domain expert, the data mining consultant will form a symbiotic relationship that provides the benefits of knowledge transfer, redundancy and inherent reinforcement training. This accumulated knowledge drives well-informed choices, validates sound judgment and combines the perspective that practically assures a solid definition of data mining project success, and then achieves it.

2. Assessment

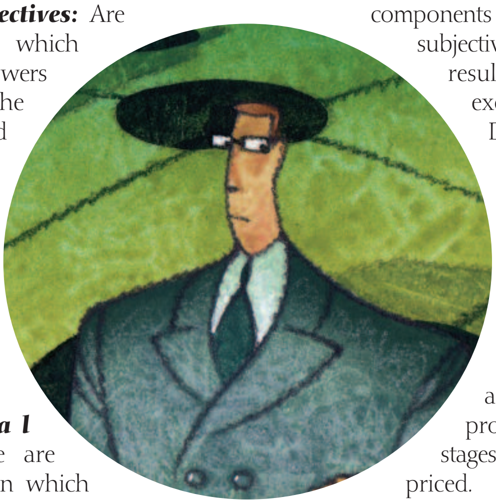
This is the stage in which the true *buy* for data mining occurs. Unfortunately, many organizations are reluctant to engage in a data mining project assessment (DMPA), because they have been burned on assessments by services companies who basically charge to exploit opportunity from their client. When done properly, however, the DMPA is an essential component of a successful data mining project.

From the client's perspective, any assessment is risky. The value of the results is unknown in advance. A full data mining implementation cannot be estimated in dollars or time prior to this exercise. There are far too many unknown factors that can dramatically affect the approach and scale of a data mining project. Further, a DMPA may reveal that an organization is not even at the starting line - thus saving substantial time and money resources by preventing a premature project. When performed by a reputable services company, this aspect of the assessment can arrive at the precise opposite of exploiting opportunity by saving needless effort and expense in advance.

The DMPA should offer a comprehensive situational report of findings that support a draft overarching plan (later described as the recommendations report). The findings report should manifest the readiness of numerous factors that need to be present for a successful data mining implementation. To name a few:

- ***Data Certification:*** A topical survey of the structure and nature of the data to support predictive analytics.
- ***Existing Resources:*** Additional tools may be recommended to support or replace existing products. Are the skills available in house to support the modeling process after deployment? What other technologies or methods have been used in the past? Are previous performance benchmarks available?
- ***Stakeholder Objectives:*** Are the questions to which executives seek answers aligned with the resources amassed in the findings? Are there desired and/or required performance levels? Are the benchmarks realistic from the consultant's experience?
- ***Functional Managers:*** There are many situations in which companies are either unable or unwilling to take the actions recommended by the model. (In the words of Jack Nicholson in *A Few Good Men*, it should be determined in advance if "You can't handle the truth!")
- ***Constraints:*** Are there hard boundaries that must be identified and built into the decision process - either before or after the model's implementation? Because virtually all data mining methods present a tradeoff between accuracy and explainability, a point on the scale should be defined. What are tolerable levels of false positives or negatives from the model?
- ***User Buy-in:*** If they won't adopt it, why build it? How may the system be designed to encourage dedicated use?
- ***IT Support:*** While usually not a deal-killer, IT is typically far more willing to support the model's function when they are included in the strategy and are invited to become data mining advocates. If IT is going to support another project that requires data access, it helps if they can also appreciate the high-level vision and benefits to the organization.

Without the DMPA exercise, some modeling projects can be carried to completion tactically, but the results ultimately do not pass the "so what" test strategically. In this situation, the client and consultant stare at each other, wondering whether they arrived at outstanding results or overall failure.



The DMPA should be created independently, allowing the organization to freely choose how the resulting plan will be developed and implemented - whether by the same services company who submitted it, another third-party vendor or the client itself. The consultant who conducts the DMPA should not incorporate proprietary components or aspects that subjectively commit the resulting build-out exclusively to the DMPA author. The value of the DMPA is all in the strategy, not the tactics.

The recommendations report from the DMPA will produce an overarching project plan. Early stages may be firmly priced. However, later stages may only be estimated because it cannot be known in advance what information will be derived from the data and how it should be leveraged. Newly discovered information can drive the remaining project in slightly different directions. Most times, there is not a significant departure from the overarching plan, but it is not realistic to ever fix the price of a data mining project beyond a few near-term tasks. This does not make data mining any easier to buy, but risk cannot be effectively managed without an adjustable, staged approach to a project that by its nature is about discovery of the unknown. A flexible structure and repeatable process with sound guidelines must be designed to effectively manage discovery.

3. Strategy

Data mining strategy is far too often overlooked or retrofitted to a resulting model. Nearly all neophytes to the data mining process are anxious to run straight to the data and push whatever is readily available into an analytical tool. While modern tools may help to some degree in data preparation, exploration and visualization, even the best tools on the market cannot anticipate, interpret or implement around environmental and political aspects of model integration. Moreover, the modeler may take poor results and proceed as though they were superior, or obtain great results (thanks to modern software's automation and wizards) and not know it - essentially building excellent models that answer the wrong questions.

Most of the strategy framework is established during the DMPA. As discoveries unfold and unforeseen information from the data is interpreted, the strategic direction may adjust somewhat, but usually not dramatically. This is why planning for a flexible framework is such a critical component for a successful data mining implementation. Any vendor who claims to have a complete framework to fit your organization's overall situation and forego a DMPA should be regarded with some suspicion. The purpose of the DMPA is to assess the overall situation and resources for data mining and draft the overarching strategy to direct the project to completion - and beyond.

4. Implementation

Thanks to automated software with effective wizards, the implementation is arguably the easiest and least risky part of a full-scale data mining project. It is far better to have a mediocre model with solid strategy than the inverse.

One misconception about selecting an external data mining consultant is that the consultant should also be an expert in your industry. It may be helpful for the consultant to have background in order to speak and interpret industry lingo while appreciating the competitive environment and primary drivers, but unlike building a knowledge base, it is actually preferable not to have the industry's strongest domain expert who also happens to do some data mining. While the consultant may appear impressive at the outset, too much industry expertise can introduce subjectivity and preconceived notions that may skew the way models are developed and interpreted.

Models by their nature are objective, and the consultant should be, too. The best results are achieved when the data mining expert drives the model-building process but not the results. The data mining consultant should then work with your organization's domain expert to mutually interpret the results, validate them and determine the most effective way to make them useful.

5. Iteration

Many industry standard and best practice process diagrams show data mining as a linear process, ending with deployment. Rather, an ongoing process toward analytic enlightenment is a more realistic expectation and effective mind-set from which to work.

As part of the assessment stage, a feedback strategy should be derived to capture valuable performance results (in marketing, this is referred to as a solicitation file). Not only is the results data used for model validation, but it is also valuable fuel for the next iteration of model building. The model should be updated and enhanced with the latest performance data, perhaps even weighting the latest feedback more heavily to encourage a greater recency effect for novel behavioral patterns.

The closing of this loop from results interpretation to model update is an excellent opportunity for reinforcement training. Reconvene with your data mining trainer or consultant to review the first pass and prepare for the next iteration. Advanced or alternative approaches for the next model update should be considered at this milestone, encouraging a progression of knowledge transfer while pushing model innovation.

While this article has presented frameworks for both failing and succeeding in data mining, the most critical phase is the data mining project assessment. All other aspects of the process are forgiving and easily repairable. Foregoing a data mining expert's comprehensive situational and goal-driven assessment is a costly way to arrive at the false conclusion that predictive analytics is overrated.

Once you have gained a base education in data mining strategy and methods and have commissioned a thorough project assessment, you will be well on your way toward data mining success. When the starting line is measured and the right framework is established, the rest of the journey is relatively straightforward. You may proceed through the discovery process with a structured yet flexible plan for nearly any scenario, confident that you will reap tremendous rewards for having taken the right approach to data mining.

Eric A. King is president and founder of The Modeling Agency (TMA). TMA is a structured team of senior consultants that provides training and consulting in predictive modeling for those who are data-rich, yet information-poor. King holds a BS in computer science from the University of Pittsburgh and has focused on data mining business development and project management since 1990. Prior to TMA, King worked for NeuralWare, a neural network tools company, and American Heuristics Corporation, an artificial intelligence consulting firm. He may be reached at eric@the-modeling-agency.com or (281) 667-4200 x210.

Editor's Note: *The Modeling Agency offers on-line, on-site, and public vendor-neutral courses in predictive modeling for practitioners and managers who are ready to implement data mining solutions. For info:* ***www.the-modeling-agency.com/dmr-special***