17

Data Mining in Business

Objectives:

- To help data engineers in a large corporation investigate the bad debts database and uncover useful patterns in selecting targets for debt recovery, thereby dramatically improving the corporation's debt recovery.
- To understand the difference between the results of the "average" practitioner and the "quality" practitioner
- To find the right balance between software, intellectual property, and so forth, is all part of the evolution of the industry.
- The business expert not only uses the results of data mining but also *evaluates* them, and this evaluation should be a continual source of *guidance* for the data mining process.
- The process must be thoroughly domain-oriented rather than technically oriented, and the tools must support an interactive, incremental, and iterative style of work.
- Data mining techniques can be implemented rapidly on existing software and hardware platforms across D&B to enhance the value of existing resources, and can be integrated with new products and systems as they are brought on-line.
- The commercial success of data mining lies in providing true value to the businessperson in a form that can be used and understood by the business community.
- Data mining tool aims to empower the business analyst to explore and understand the dataset in relation to his/her own knowledge, rather than aiming to replace the analyst with some automated data-discovery algorithm.

Introduction

In this section, data mining for business problems, business focus on data engineering, overview on business applications, and data mining in business case studies are presented.

S. Sumathi and S.N. Sivanandam: *Data Mining in Business*, Studies in Computational Intelligence (SCI) **29**, 387–409 (2006)
www.springerlink.com

Data engineering is inherently a multidisciplinary field, because of the number of technologies involved: visualization, data analysis, knowledge engineering, perhaps databases, and courses the subject matter of the application. So there we have the technical background of the community, and some idea of the range of applications. What are the business implications here? A number of factors have emerged in the consulting works that are beginning to give a better understanding of the business nature of the discipline. First, the community has a number of different focuses. The business expert plays a critical role in data mining, both as an essential source of input (business knowledge) and as the consumer of the results of data mining. Data mining can reveal patterns in data, but only the business expert can judge their usefulness. It is important to remember that the data is not the business, but only a dim reflection of it. This gap, between the data and the business reality it represents, we call "the chasm of representation" to emphasize the effort needed to cross it. Also this section discusses a case study on "*An Overview of Data mining at Dun & Bradstreet (D&B)*".

17.1 Business Focus on Data Engineering

Data mining has emerged this decade as a key technology for areas such as business intelligence, marketing, and so forth. For the purposes of discussion, application, and business domains we consider here include telecommunications, medical devices, space science (vehicle health management and scientific instrumentation), targeted marketing, and mining.

From a technical view, we do not consider data mining to be a new field, but rather another discipline in the lengthy history of engineering sciences that use data is a core focus for developing knowledge. This family of disciplines we consider here comes under the term *data engineering*. Some traditional and nontraditional examples follow: Data engineers work with physicists in analyzing spectral data measured from a high-resolution imaging spectrometer develop sophisticated models of the spectrometer's complex error modalities (registration, response function, calibration, measurement glitches) so that a high-fidelity model of the spectrometer's measurements can be developed. Data engineers investigating the performance of an industrial strength place-and-route package uncover useful characteristics of the optimization process and thereby improve the performance of the algorithm. Data engineers work with astronomers in analyzing infrared data from an electronic star catalog. The analysis, in concert with the astronomer's interpretations reveals new, publishable classes of stars and also uncovers troublesome, never-before recognized artifacts with the original instrument. Data engineers in a large corporation investigate the bad debts database and uncover useful patterns in selecting targets for debt recovery, thereby dramatically improving the corporation's debt recovery.

At the time of the development, the individuals performing these tasks may have considered themselves applied machine learning researchers, decision analysts, statisticians, or neural network researchers; however they were all performing data engineering. We may have also head of the terms data mining and knowledge discovery, exploratory data analysis, intelligent data analysis, and so forth. These areas performing similar tasks however have a particular emphasis that distinguishes their origins, whether it be the applications they serve of the algorithms for data analysis that they use.

First, the community has a number of different focuses.

- First, there are more common turnkey applications. A well-known example is HNCs pushing into credit applications. We do not consider this data mining, but rather an application with a small component of data mining technology embedded.
- Second, there are general-purpose tools. Like the tools market everywhere, and as Kohavi (SGI – Silicon Graphics Institute) has said in his recent talks, this is a consolidating market.
- Then there are special purpose developments, which involve the technology we have referred to above as data engineering.

The experiences in this third focus period an interesting conundrum for the business manager. We find that in this third focus, there is a big difference between the results of the "average" practitioner and the "quality" practitioner. Every good software manager would know that a really good programmer can produce 100 times more code than an average programmer, partly due to the net result of subsequent maintenance, reduction in overhead and systems validation, and so forth. We find the same with data engineering. Except with data engineering, we find there are few key insights made in a project that make all the difference. Mundane use of the "usual tools" in the "usual manner" by the average practitioner gets one so far. But a big difference in performance is gained by the quality practitioner who makes a few key insights to change the project.

What are the business implications here? There are several.

- We often find businesses are reluctant to put outside specialists onto key technology development regarding perceived ownership of their intellectual property. If there core business involves software, they do not like to be seen as needing outside consultants developing part of their proprietary software. Network management companies are more reluctant to seek outside consultants, than say, medical devices companies. Because advances in these areas often need some key technology insights, the development in the network management company simply remains untouched unless they have an in-house, experienced, research and development group knowledgeable with the right methods (such as Microsoft, HP, etc.) We see this as damaging to the application domains themselves. Most importantly, we believe this effect favors the multinationals over the small businesses,

usually considered the engine of innovation. This is a problem with the business handling of intellectual property.

- Second, we fear, but have not yet seen a backlash against data mining as had occurred for expert systems. Large corporations with ideal and potentially lucrative applications throw inexperienced staff together with the usual tools in poorly conceived projects. The broad experience in applications has shown that except in circumscribed vertical markets where complete systems can be developed by professionals the key insights required to make the difference can require the right professionals on the task. Thus many may suffer. We have anecdotal evidence here from data mining consultants. From a distance, business managers would blame the technology and not their implementation.
- Third, we are developing a software environment where we can address the kinds of developments we believe are necessary in those domains where specialist R & D is needed. MindSet from SGI, for instance, is successful because it provides an easy-to-use but powerful environment for visualizing data. Tools such as SAS rely on the large number of SAS-trained users in industry, for instance prior to the growth in data mining. The in-house focus for supporting consulting is to provide common platforms and tools, and plug compatible algorithms with modular components so that we can efficiently customize systems for the clients to focus on their needs.

Regardless of what happens to data mining as a community, we know that data engineering in one form or another will continue to remain a key enabling technology for many businesses, and thus finding the right balance between software, intellectual property, and so forth, is all part of the evolution of the industry.

17.2 Data Mining for Business Problems

As data mining and machine learning techniques are moving from research algorithms to business applications, it is becoming obvious that the acceptance of data mining systems into practical business problems relies heavily on the integration of the data mining system in the business process. Some key dimensions that data mining developers must address include understanding the business process from the end user perspective, understanding the environment in which the system will be applied, including end users throughout the lifecycle of the development process, and building user confidence and familiarity of the techniques.

One critical aspect of building a practical and useful system is showing that the techniques can tackle the business problem. Traditionally, machine learning and data mining research areas have used classification accuracy in some form to show that the techniques can predict better than chance. While this is necessary, it is not sufficient to sell data mining systems. The evaluation

methods needs to more closely resemble how the system will work while in place.

One way to more closely evaluate data mining software in their intended setting is to incorporate time into the evaluation process. Although the time issue makes the prediction scenario more complicated, many data warehouses have data that is time dependent. For example, billing data stores billing, payment, and usage data for each customer indexed by time. As Kurt mentions in "Some thoughts on the current state of data mining software applications," few if any data mining systems deal with the time variable indirectly stored in the data.

Although the time dimension is left out during the prediction process, data mining systems should be evaluated with some aspects of time kept in mind. Researchers and developers can simulate time-dependent evaluation by evaluating models on historical data stored in the data warehouse. For example, suppose we are building a model to predict whether a customer will churn in a given month. Suppose we have the data for the independent variables at time t and we make a prediction for person x saying that they will churn. When will x churn? Will they churn immediately, in the next two weeks, in the next month? For what time period should we evaluate the model? An intuitive guess would say that the model is most accurate at predicting churn the closer it occurs in time to the independent data. When does the model predict the same as the background churn rate? While accuracy is a valid evaluation criterion, determining how long the model is valid is also important information. Instead of showing the accuracy of a model as a single number, the accuracy could be shown as a function of time. This information can also be used to compare different data mining techniques. The characteristics of the model should be tested while increasing time to better simulate the data mining software while used in a business process. The accuracy of the model in given time and other evaluation criteria should be provided to the end users, so they can determine those characteristics that are more important to their business task.

Given that models have some accuracy function implies that models should be relearned or refreshed after some amount of time. If the accuracy of a model comes close to random chance after some time, a new model should be learned. The older model and the newer model should be somewhat consistent and similar. For instance, if a model at time t is based on attributes A, B, and C, we would expect the refreshed model to use a similar set of attributes. If they change radically, the models may be overfitting the data or the models may reflect seasonal trends. The end users of the system expect the models to be somewhat consistent. They might lose confidence if the models change radically, because intuitively the radical change may not make sense.

As stated earlier, the evaluation should reflect the business process it will be applied in. For example, if the churn system is used to identify churners monthly for targeted campaigns, then an interesting question from an end user may be to ask what percentage of churners in month y would the data mining

software predict to churn ahead of time. The results of running experiments on historical data to answer this question may give some indication of how often campaigns need to be run to capture a certain percentage of churners. In addition, by noting those customers that end up being on the predicted churn list for successive months we also find out more about the consistency of the models. These types of questions come about by interacting with end users and by looking at the task through their perspective.

Unique Perspective on Data Mining Problems

The combined experience in applying data mining technology to many domains over the years has taught several lessons that are not commonly discussed in the community, by either vendors, researchers, or business users. The important three problems are presented as follows:

1). Before business problems can be solved with data mining, they must be transformed to match existing tools.

Data mining tools perform a small set of basic tasks such as classification, regression, and time-series analysis. Rarely is a business problem exactly in one of these forms. Usually it must be transformed into (or rephrased as) one of these basic tasks before data mining tool can be applied. Often, in order to solve a problem it must be decomposed into a series of basic tasks. Indeed, much of the art of data mining involves the creative decomposition of a problem into a sequence of such subtasks that are solvable by existing tools.

For example, the work on cellular phone fraud detection transformed the problem of fraud detection into a sequence of knowledge discovery, regression, and classification tasks (mining for indicators of fraud, profiling customer behavior, combining evidence to classify behavior as fraudulent). No single type of task was adequate to solve the problem.

2). Evaluation of data mining results is more complex than either developers or users believe.

Most data mining tools, like the research prototypes from which they were derived, measure performance in terms of accuracy or classification error. A tacit assumption in the use of classification accuracy as an evaluation metric is that the class distribution among examples is constant and relatively balanced. In the real world this is rarely the case. Classifiers are often used to sift through a large population of normal or uninteresting entities in order to find a relatively small number of unusual ones; for example, looking for fraudulent transactions or checking an assembly line for defective parts. Because the unusual or interesting class is rare within the classification is rare within the general population, the class distribution is very skewed.

Evaluation by classification accuracy also assumes equal error costs. In the real world this is unrealistic because classifications lead to actions, which have consequences, sometimes grave. Rarely are mistakes evenly weighted in their cost. We have yet to encounter a domain in which they are. The class skew (as well as error costs) may change over time, after a data mining solution is deployed. Indeed, error costs and class distribution in the field may never be known exactly.

Unfortunately, the importance and difficulty of evaluation is often not appreciated by business users either. The business user usually knows the general problem to be solved, but may not be able to specify error costs or even advise in their calculation. Sometimes the business user does not know how well current procedures solve the problem, and has no mechanisms in place to evaluate their performance. We are sympathetic to this, since evaluating performance often tasks time and effort away from the task itself. However, it makes measuring the efficacy of a data mining solution difficult or impossible. These recurring difficulties with evaluation have directed the research at Science & Technology Center Bell Atlantic. They have developed a technique based on ROC analysis that greatly facilitates comparison and evaluation of data mining results. The technique is especially useful when error costs and class distribution are only known approximately, or may change.

3). Data preparation and data cleaning are more time consuming and knowledge intensive than is acknowledged.

In our experience, understanding the data, reducing noise, and converting the data to an appropriate representation is the most time-consuming part of the data mining process. Furthermore, the process is usually iterative and knowledge intensive: as the project progresses, we learn more about the process that generates the data and we have to go back and re-clean them based on the new knowledge. Although the provider usually has information about the data, we are often the first people ever to analyze the data carefully. We have uncovered errors, idiosyncrasies, and artifacts of the data gathering process that were unknown to the provider. These discoveries sometimes end up changing how we approached the data mining task. Data preparation and cleaning are often tedious, uninteresting tasks. However, over the life of a data mining project, these tasks account for far more time than the taken by applying the machine learning algorithms.

Benefits of a Standard Data Mining Process Model

Data mining success stories have triggered increased interest within the business community, particularly in large corporations with vast stores of data about their customers and business operations. Their interest appears to be following a path similar to that of early research in machine learning: the tendency to view data mining as the isolated application of a data mining

algorithm to a pre-existing dataset, where the key determinant of success is selecting (or creating) the "best" model-building algorithm.

As businesses continue to use data mining technology, they are likely to discover, as experienced practitioners and researchers already have, that:

- There is usually little difference in accuracy between modeling algorithms.
- Availability of useful data, dataset preparation, and user skill are more important than which algorithm is chosen.
- Model development is more properly viewed as a multistep *process*, of which application of the modeling algorithm is only a small part (practitioner's informal estimates tend to range from about 10% to 30% of the total effort).

While business customers would eventually learn these lessons through experience, it is hazardous for the health of the industry to allow this tool-centric focus to continue. At this early stage, too many disaster stories would be fatal to the field. Tool vendors have found it effective to sell tools based on the premise that they are like golden geese – feed data at one end and golden nuggets of knowledge will magically come out the other. However, when someone's goose produces the wrong kind of nuggets, their first impulse is not to question whether they are properly skilled in raising golden geese or providing the right food. They will scream "it's a hoax – data mining doesn't work." If so, data mining will be discredited like other technologies that were overhyped and failed to meet the inflated expectations.

For the data mining industry to prosper on a wide scale, it is necessary to create the perception that "data mining works." It is fine if the perception is "data mining works – if done properly." Of course, then the obvious question is "how?" For acceptance to spread beyond the "early adopters," customers must feel confident that they will know how to manage a data mining project to ensure success. A major part of the comfort level is to understand what the stages of a project are, what issues will be need to be dealt with, what tasks will need to be performed, etc. Having a process model goes a long way toward creating this initial comfort level. Customers would not feel as if they are wandering into completely uncharged territory. They realize they will encounter many difficult situations that will need to be dealt with, but at least they known what they are likely to encounter. A service provider that can communicate a convincing process will have a big advantage over one that cannot.

The existence of a process model is a big improvement over the situation without one. However, a great deal of confusion will result if a customer is presented with several different process models. It might be that in reality they are very similar. However, they might sound quite different to a customer who is not a data mining expert, and is then presented with the dilemma of trying to decide if one is better then the other. Faced with such a situation, the conservative customer may well delay data mining project, preferring to wait until the picture becomes clearer.

Contrast this with the situation where nearly even prospective service provider describes the same, standard, data mining process model. This removes a major obstacle to the decision to do a project. The customer is still confronted with which service provider to hire. But he/she will no longer need to get tangled up in confusing arguments about which process models are really different and which is best. The key issue is that if a sufficient comfort level is achieved, the customer will be much more willing to proceed with the project.

Benefits of a Common Process Model

By adopting and promoting a common view of the data mining process, the data mining community would benefit in a number of ways:

- A standard would increase the comfort level of potential customers of data mining products and service.
- A greater proportion of efforts would be successful, because they followed established principles, increasing the perception of data mining as an effective technology and a high-return investment.
- Projects are more likely to proceed smoothly if all parties understand the normal course of data mining project.
- A standard process improves communication about data mining topics by providing a common reference point for customers and clients to compare products and services.

However, different groups are affected differently by standard model. The major communities, and their expected relation to a process model, are discussed below.

Tool Vendors

For tool vendors the key question is the market's verdict of the value of data mining. If the perception is positive, they will buy lots of tools. Of course, some vendors will succeed more than others. Naturally, the prospective market for tools is significantly larger if the belief of the market is that data mining can be successfully performed by less experienced people. However, if the verdict is that data mining is another overhyped hoax, the market will shrink, and no one will sell many tools. In this case, tool vendors will suffer the fate of any company whose habitat disappears: most will become extinct, others will somehow find a way to adapt to a new environment. The conflict for tool vendors is that there is a short-term benefit to downplaying the amount of effort and skill needed to achieve high-quality results.

Service Providers

Service providers have a different conflict. Certainly, they would benefit if the perception becomes "Data mining can produce great rewards, but only if performed by experts who know what they're doing," i.e., data mining service providers. In this situation, a service provider with a high-quality, proprietary process has a distinct competitive advantage. However, prospective customers confronted with arguments over conflicting process descriptions might decide to stay out of the market. Adoption of a common view of the data mining process should increase the total market for data mining services, though it might be harder to differentiate offering. In such a world, a propriety process model could turn from an asset into a liability; prospective customers will question why your process is different form the industry consensus. It will be up to the service provider to justify excluding a standard task or to explain the added benefit of additional tasks.

End Users

Less experienced users would probably be most eager to embrace a process model, for the guidance it would provide. At first glance, experienced modelers might view a standard process model as a threat, because it would provide greater benefit to less experienced practitioners. However, the greater the demand for data mining services overall, the greater will be the demand for experts. In addition, the existence of a widely known model should make it much easier for the analyst to communicate what they are doing to a client. In addition, the client is much less likely to question about the necessity of tasks that are described in a standard process.

Business Customers

Probably the most enthusiastic advocates of a standard process model. These are the poor folks who, despite limited technical understanding of data mining, must somehow sift through conflicting definitions and marketing claims and decide how their organization will use data mining. A common version of the data mining process provides them with a framework for structuring their projects and for evaluating tool and service offerings.

17.3 Data Mining and Business Intelligence

Data mining is about finding *useful* patterns in data. This word *useful* can be unpacked to expose many of the key properties of successful data mining.

The patterns discovered by data mining are useful because they extend existing business knowledge in useful ways. But new business knowledge is not created "in a vacuum"; it builds on existing business knowledge, and this

existing knowledge is in the mind of the business expert. The business expert therefore plays a critical role in data mining, both as an essential source of input (business knowledge) and as the consumer of the results of data mining.

The business expert not only uses the results of data mining but also *evaluates* them, and this evaluation should be a continual source of *guidance* for the data mining process. Data mining can reveal patterns in data, but only the business expert can judge their usefulness. It is important to remember that the data is not the business, but only a dim reflection of it. This gap, between the data and the business reality it represents, is called *the chasm of representation* to emphasize the effort needed to cross it.

Patterns found in the data may fail to be useful for many different reasons. They may reflect properties of the data, which do not represent reality at all, for example when an artifact of data collection, such as the time a snapshot is taken, distorts its reflection of the business. Alternatively, the patterns found may be true reflections of the business, but they merely describe the problem that data mining was intended to solve – for example arriving at the conclusion that "purchasers of this product have high incomes" in a project to market the product to a broader range of income groups. Finally, patterns may be a true and pertinent reflection of the business, but nevertheless merely repeat "truisms" about the business, already well known to those within it.

It is all too easy for data mining, which is insufficiently informed by business knowledge to produce useless results for reasons like the above. To prevent this, the business expert must be at the very heart of the data mining process, spotting "false starts" before they consume significant effort. The expert must either literally "sit with" the data miner, or actually perform the data mining. In either case, the close involvement of the business expert has far-reaching consequences for the field of data mining.

Data Mining as a Nontechnical Process

Business experts are seldom also technical experts, and their deep involvement in data mining has a fundamental effect on its character. The process of data mining is one in which the business expert interprets the data, a simple extension of ordinary learning by experience. In such a framework, technology must as far as possible remain hidden, while revealing the patterns in the data.

The organizers rightly state that successful business data mining does not come down to "hot algorithms." Equally irrelevant to the core process of data mining are database support, application integration, business templates, and scalability; data mining tools may usefully have such attributes but they are essentially *technological* properties. The business user must be able to approach data mining as a window on the business and engage *with the data* without the distraction of technological detail.

Data Mining Tools

The requirement of data mining to be accessible to business experts also shapes the requirements for data mining tools. These end-user-oriented requirements can be described in many different ways, but here we focus on three key properties: data mining tools must be *interactive, incremental*, and *iterative.*

Interactive: Modern "desktop" applications are highly interactive as a matter of course, but here we focus on a deeper interpretation of "interactive": the user must be enabled to *interact with the data*, and not just with the technology. The user interface of data mining tools should be designed to highlight the properties of data and play down the details of technology, whether that technology be database links, efficient indexing, visualization display parameters, or machine learning algorithm.

Incremental: The data mining process is incremental, each successive investigation builds on the results of the previous one; thus the principle learning from experience applies not just to the data mining exercise as a whole but also to each step within it. Data mining tools must be designed to encourage this re-use of results as the data miner, in a step-by-step manner, *builds up a picture* of the patterns in the data. This means that data mining tools must be highly integrated; query must lead naturally to visualization, visualization to data transformation and modeling, and modeling to visualization or further queries. These transitions are merely examples; overall the process must appear seamless, with the effective methods of investigation at any point being also the most obvious, and without the intervention of technological barriers or distraction.

Iterative: Data mining is seldom a simple linear process; successive steps not only build on one another's results, but also refine the approach of earlier steps. For example the results of modeling may show that the data should be further refined and the modeling repeated, or may point to areas for closer examination in an earlier data exploration phase. Any result may point to earlier steps, refining not only the data but also the process itself; each step also has the potential to open up entirely new avenues of inquiry. It should be emphasized that the process is not organized into discrete steps concerned with different types of knowledge; rather the discovery of detailed properties of the data proceeds alongside a gradual refinement of the business concept involved, and the unfolding of key patterns to be utilized.

The iterative nature of data mining is apparent at a variety of levels. For example at the detailed level, a modeling process may be repeated many items (and gradually transformed), for example in the space of a day. Many models are built over this time, and each contributes a small "nugget" of knowledge to the overall process, we might call these "throw-away" models – they are formed to be read, digested, and then thrown away. At the overall project level, the data mining process is also iterative, and should for a project

of significant duration contain "planned in" iterations for the production of improved models or other results.

Data mining tools must be designed to support this iterative property of the data mining process. The requirements here are similar but not identical to those relating to the "incremental" property. Data mining operations, once configured, must not be "set in stone" – they should be *designed to be refined* in the light of subsequent events.

17.4 Data Mining in Business - Case Studies

Case Study 1 – An Overview of Data mining at Dun & Bradstreet (D&B): This case study is taken from Data Intelligence Group (DIG), D&B, 1995.

Executive Summary

Data mining, the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help (D&B) "preemptively define the information market of tomorrow." Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by retrospective tools. Data mining answers business questions that traditionally were time consuming to resolve. Data mining tools scour databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations.

D&B companies already know how to collect and refine massive quantities of data to deliver relevant and actionable business information. In this sense, D&B has been "mining" data for years. Today, some D&B units are already using data mining technology to deliver new kinds of answers that rank high in the business value chain because they directly fuel return-on-investment decisions.

Data mining techniques can be implemented rapidly on existing software and hardware platforms across D&B to enhance the value of existing resources, and can be integrated with new products and systems as they are brought on-line. When implemented on high-performance client-server or parallel processing computers, data mining tools can analyze massive databases while a customer or analyst takes a coffee break, then deliver answers to questions such as, "Which clients are most likely to respond to the next promotional mailing, and why?"

In the D&B units DIG surveyed, we found strong interest and a wide range of activities and research in data mining. Groups are engaged in data mining to varying degrees, from experimentation by individual analysts to

the deployment of completed projects. We also found a wealth of potential business opportunities that could open up through data mining technology.

The breadth of D&B's collected data places the company in a unique position to take advantage of the fact that data mining tools produce better results with larger, broader databases. By integrating data mining into its products and services, D&B can leverage its exiting resources to achieve new revenue.

A Strategic opportunity for D&B

D&B units are connected through the common goal of delivering integrated, global solutions to support business decisions. In accomplishing this goal across a broad spectrum of markets, D&B units face similar market pressures and opportunities. For example, customers urgently require tools to help them keep pace with accelerating growth in the size and complexity of business data. At the same time, customers demand ever more timely, sophisticated, and widely integrated data analyses.

D&B units work hard today to maintain their leadership against a growing competitive threat from other vendors. These vendors have often aggressively exploited new technology to capture market advantage. While D&B units have responded successfully to these competitive challenges in the past, the question remains: What new technology is becoming available today the D&B can leverage proactively?

Data mining is such a technology. D&B is in a unique position to take the lead in delivering the benefits of data mining technology to customers. The company has a wealth of data unrivaled in its breadth and depth, and the understanding of the relevant markets that is necessary to bring this technology to customers successfully. D&B units are engaged in markets where data mining can have significant impact. These markets use large databases and need the power of data mining to achieve a better understanding of their data.

The Scope of Data Mining

Data mining derives its name from the similarities between searching for valuable business information in a large database – for example, finding linked products in Nielsen's gigabytes of store scanner data – and mining a mountain for a vein of valuable ore. Both processes require either sifting through an immense amount of material, or intelligently probing it to find exactly where the value resides.

Given databases of sufficient size and quality, data mining technology can generate new business opportunities by providing these capabilities:

- ***Automatic prediction of trends and behaviors.*** Data mining automates the process of finding predictive information in large databases.

Questions that traditionally required extensive hands-on analysis can now be answered directly from the data – quickly. A typical example of a predictive problem is targeted marketing. Data mining uses data on past promotional mailings to identify the targets most likely to maximize return on investment in future mailings. Other predictive problems include forecasting bankruptcy and other forms of default, and identifying segments of a population likely to respond similarly to given events.

- ***Automatic discovery of previously unknown patterns.*** Data mining tools sweep through databases and identify previously hidden patterns in one step. An example of pattern discovery is the analysis of retail sales data to identify seemingly unrelated products that are often purchased together. Other pattern discovery problems include detecting fraudulent credit card transaction and identifying anomalous data that could represent data entry keying errors.

Data mining techniques can yield the benefits of automation when implemented on existing software and hardware platforms at D&B, and can be implemented on new systems, as existing platforms are upgraded and new products developed. When data mining tools are implemented on high-performance parallel processing systems, they can analyze massive databases in minutes. Faster processing means that users can automatically experiment with more models to understand complex data. High speed makes it practical for users to analyze huge quantities of data. Larger databases, in turn, yield improved predictions. Databases can be larger in two senses:

- ***Higher dimensionality.*** In hands-on analyses, analysts must often limit the number of variables they examine because of time constraints. Yet variables that are discarded because they seem unimportant may carry information about unknown patterns. High-performance data mining allows users to explore the full dimensionality of a database, without pre selecting a subset of variables.
- ***Larger samples.*** Larger samples yield lower estimation errors and variance, and allow users to make inferences about small segments of a population.

A recent Gartner Group Advanced Technology Research Note listed data mining and artificial intelligence at the top of the five key technology areas that "will clearly have a major impact across a wise range of industries within the next three to five years." Gartner also listed parallel architectures and data mining as two of the top ten new technologies in which companies will invest during the next five years. According to a recent Gartner HPC Research Note, "With the rapid advance in data capture, transmission and storage, large-systems users will increasingly need to implement new and innovative ways to mine the after-market value of their vast stores of detail data, employing MPP [massively parallel processing] systems to create new sources of business advantage (0.9 probability)".

At D&B, data mining technology provides a basis for new products and for enhancements to existing offerings. For example, at DBIS, data mining tools can be used to automate more elements of the process of building risk models for a variety of markets. Data mining can present a Nielsen customer with the top ten most significant new buying patterns each week, or present an IMS customer with patterns of sales calls and marketing promotions that have significant impact within certain market niches.

The most commonly used techniques in data mining are:

- ***Artificial neural networks:*** Nonlinear predictive models that learn through training and resemble biological neural networks in structure.
- ***Decision trees:*** Tree-shaped structures that represent sets of decisions. These decisions generate rules for the classification of a dataset. Specific decision tree methods include classification and regression trees [CART] and chi square automatic interaction detection (CHAID).
- ***Genetic algorithms:*** Optimization techniques that use processes such as genetic combination, mutation, and natural selection in a design based on the concepts of evolution.
- ***Nearest neighbor method:*** A technique that classifies each record in a dataset based on a combination of the classes of the k record(s) most similar to it in a historical dataset (where k > 1). Sometimes called the *k-nearest neighbor technique.*
- ***Rule induction:*** The extraction of useful if-then rules from data based on statistical significance.
- ***Data Visualization:*** The visual interpretation of complex relationships in multidimensional data.

Many of these technologies have been in use for more than a decade in specialized analysis tools that work with relatively small volumes of data. These capabilities are now evolving to integrate directly with industry-standard data warehouse and OLAP platforms.

The Foundations of Data Mining

Data mining techniques are the result of a long process of research and product development. This evolution began when business data was first stored on computers, continued with improvements in data access, and more recently, generated technologies that allow users to navigate through their data in real time. Data mining takes this evolutionary process beyond retrospective data access and navigation to prospective and proactive information delivery.

Data mining is ready for application in the business community because it is supported by three technologies that are now sufficiently mature:

Massive data collection and storage
Powerful multiprocessor computers
Data mining algorithms

Commercial databases are growing at unprecedented rates. The accompanying need for improved computational engines can now be met in a cost-effective manner with parallel multiprocessor computer technology. Data mining algorithms embody techniques that have existed for at least ten years, but have only recently been implemented as mature, reliable, understandable tools that consistently outperform older statistical methods.

In the evolution from business data to business information, each new step has built upon the previous ones. For example, dynamic data access is critical for drill-through in data navigation applications, and the ability to store large databases is critical to data mining.

This survey of a limited number of D&B units points to quantifiable business benefits achievable through the integration of data mining technology with D&B products and services. Data mining is a powerful technology with great potential for adding value to the offerings of D&B units. D&B is in a remarkable position to take advantage of this technology: it has the data, and it has the infrastructure to support units in collaborating to solve shared problems.

Case Study 2: Success of Data Mining in Business-Data Cruncher Design Aspects - This case study is taken from Data Mind Corporation, San Mateo, CA.
Awesome data mining tools, fantastic algorithms, rapidly converging neural networks, highly accurate classification methods, clustering methodologies, etc. are neat and useful tools for the knowledge discovery professional, but they are far from demonstrating significant value to the business person. The key to commercial success of data mining lies in providing true value to the businessperson in a form that can be used and understood by the business community. We present here several of the most important aspects of how the data cruncher was designed to accomplish the business users goals.

Recognize who the customer is

To be commercially successful the first thing to realize is that the customer is not the fellow scientist. The customer is not the statistical analyst and the customer is not the mathematician. Sure, we can sell tools to all of them, but in order to be a commercial success we have to sell to the businessperson. Then we need to realize that business people spend their money in tools that help to solve specific business problems. So the tools need to demonstrate that they are useful in business situations and that they have a visible impact in the business.

Speak Business

The tools have to make themselves comprehensible to the business users. The language used has to be simple and business oriented. Results should

be explained in terms that are comprehensible to the business user. It is important to note that this is not just replacing a statistical concept by the English sentence that describes it. Rather it is realizing the communication with the user in the terms and concepts that are familiar to the user.

Task and Goal Oriented

The business user does not want to create a neural network, a decision tree, or an agent network. The user wants to solve a specific problem, to find a specific answer. The user wants to know into what segments it makes sense to divide the customers. The user wants to know what customers are most likely to churn. The user want to know how likely a specific customer will respond to a product being currently promoted.

The "data mining tools" user interface should reflect the problems that the user is trying to solve. The specific approach used in the data cruncher to attack this issue is the concept of assistants. Assistants somewhat resemble wizards in the sense that they guide the user through a set of steps, but they are more complete than wizards. They provide random access to the steps and they follow the user through the whole process, always being there, accessible, and well documented. They guide the user through the mining process. The assistants also allow the user to step outside the assistant and do things using the full flexibility of the tools when necessary, and then go back to the assistant. The data cruncher assistants are customizable to many different business situations through the use of a scripting language.

Bridge the gap between analysis and deployment

Data mining models are developed for a purpose. The data mining tools should help in allowing the user to apply the model to its purpose. For example, if the model is developed with the goal of identifying the best customers for a mailing campaign, then the model should be available where the mailing lit is built. Approaches to solve this issue include providing APIs that enable other applications to make use of the model or adding capabilities to build a mailing list to the data mining tool itself, or integrating the data mining pieces into a mailing list generation.

Another approach is to provide access to the data mining models through a service-oriented interface, where the models are published to a centralized server and then can be used by any application wanting to evaluate specific models against specific records. For example, several models maybe built to determine customer segmentation, likelihood to churn, customer value, cross-selling opportunities, etc. Then these models can be made available through the server and any number of applications can apply the models to different customers by sending the appropriate messages to the server. For example, a mailing list building application may consult a model to score the likelihood of the customers to respond to the mailer.

Combine explicit with implicit knowledge

An important aspect of bridging the deployment gap is to understand that data mining models alone cannot take decisions. The models represent the implicit or learned knowledge. Model results have to be filtered through business rules – which represents the explicit knowledge – before they are put to work. These business rules may contain overrides, additional targeting criteria, geographic or time restriction, etc. For example, a company that sells video cassettes may want to avoid offering a rated R film to a customer who is a minor, even if the cross selling model says that the customer's profile indicates that this is a good title to offer. Another example maybe targeting a churn avoidance campaign to the residents of California, in this case even if the data mining model may indicate that a customer is about to churn, the offer should not be made because the customer does not live in California.

An additional advantage of having business rules combined with the data mining models is that the same rules and models can be used at the many different points where a decision is made. For example, a marketing campaign targeted at attracting new customers may use several models, like customer segmentation, value, and likelihood to accept the offer, combined with some business logic maybe used to decide whether the offer should be made. If these models are deployed to a central server together with the business logic, then the same selection criteria can be used at the many points of contact between the company and the customer. For example, the company's call center, the mailing of the next bill, or the customer's visit to the company's Web pages.

Make it responsive and easy to change

Business situations change very rapidly. It is very important for the business user to react quickly to changing business conditions. In today's competitive world it is not acceptable to have the answer to a question be delivered three months after it was asked. For example, a marketing person developing a promotion maybe interested in modeling the customer's behavior to finetune the targeting. It is important for this model to be available very soon. In addition, the whole package that includes the several data mining models and business logic should be readily available and easy to modify to adapt to the necessary changes in the promotion.

Decision Delivery Systems

Decision delivery systems are designed as a vehicle for bringing decisions to different applications. These systems can typically combine the results of different data mining models with business logic to generate the decisions. As a centralized facility they provide a focused point for the deployment of models and knowledge, helping bridge the gap between the development of useful data mining models and putting them to work.

Business Applications- Overview

Many businesses are interested in data mining because of the falling cost of data storage, the increasing ease of collecting data over networks, and the immense computational power available at low prices. The development of robust and efficient data mining algorithms has caused most businesses to create huge databases containing as much information about their activities as possible. Already available on the market are generic multitask data mining tools to perform a variety of discovery operations. Examples include Clementine, IMACS, MLC++, MOBAL, and Recon.

Making data mining programs useful to businesses requires several elements. First, the problem needs to be started in the business users' terms, including viewing the data in a business model perspective. Second, the program needs to support specific key business analyses such as segmentation, which is very important in marketing application. Third, the results of the data mining need to be presented in a form geared to the business problem being solved. Finally, there has to be a support for protracted data mining on an increasing data set, since business databases are continually growing to store increasing numbers of business transactions.

Data mining applications have been developed for a variety of businesses, including marketing, finance, banking, manufacturing, and telecommunications. Data mining in marketing falls into the broad area called *database marketing*. It consists of analysis of consumer databases to select the best potential customers for a particular product. *Business Weekly* estimated that more than 50% of all U.S. retailers use or plan to use database marketing. American Express has had good results from database marketing, experiencing a 10% to 15% increase in credit card use.

The BBC of the U.K. hired Integral Solutions Ltd. developed a system for predicting the size of television audiences. Integral Solutions Ltd program used neural networks and rule induction to determine the factors playing the most important roles in relating the size of a program's audience to its scheduling slot. The final version performed as well as human experts but adapted more quickly to changes because it was constantly retrained with current data.

Early developments in data mining included Cover story and Spotlight, programs that analyzed supermarket sales data and generated reports on the most significant changes in volume and share broken down by region, product type, and other qualities. Causal factors like price changes and distribution channels were analyzed and related to changes in volume and share. Spotlight later grew into the opportunity explorer system, which includes support for sales representatives of consumer packaged-goods companies to examine their business with individual retailers. This is accomplished by creating presentations showing the advantages of stocking additional products or having special promotions. It even generates interactive reports using hyperlinks for easy navigation.

The Management Discovery Tool (MDT) developed by AT&T and NCR Corporation, incorporates a set of business rules so that users can easily set up monitors for detecting significant changes in important business indicators. MDT also allows automatic HTML report generation, making it easier for users to understand the causes of changes while allowing deeper analysis through point and click links. To appeal more to mainstream business users, MDT provides a limited set of analysis types, including summarization, trend analysis, change analysis, and measure and segment comparison.

The Fidelity Stock Selector fund uses a neural network data mining system to select investments. It performed quite well in the stock market overall. A human fund manager evaluates the output of the system before the action is taken however, so it is not ascertainable how to divide the credit between human and machine.

A data mining system developed by Carlberg & Associates using neural networks was used to predict the Standard & Poor's 500 index. It incorporated interest rates, earnings, dividends, the dollar index, and oil prices in its analysis. The system was amazingly successful, accounting for 96% of the variation in the Index from 1986 to 1995.

The Clone detector system developed by GTE Corporation uses customer profiles to detect cellular cloning fraud. If a particular user suddenly starts calling in an unusual way, the Clone detector automatically informs GTE security. A similar system was developed by AT&T to detect international calling fraud, but that system is much more interactive with and reliant upon human operators.

AcknoSoft developed a data mining program called CASSIOPEE for General Electric and SNECMA. CASSIOPEE is being used by three European airlines to diagnose and predict technical problems in Boeing 737 aircraft. Clustering methods are used to derive families of faults.

Not All In the Data

This case study is taken from QuadStone Ltd, Edinburgh, Scotland. Whatever be the various competing definitions of "data mining" used, it is clear that it has something to do with finding patterns in data. There would also probably be fairly broad agreement where desirable properties of patterns found would include

- Meaningful
- Comprehensible
- Expressible
- Actionable

It is important to distinguish between various different kinds of information that can be relevant to data mining. Perhaps the most important of these might be classed as:

1. Information contained in the dataset being mined in a form that is reasonably accessible to automatic pattern detection, whether with traditional statistical methods, machine learning methods, or other automated procedures (e.g., a strong, meaningful correlation between one or more "independent" variables that exist as fields in the data and a dependent variable being modeled that also exists as a field in the data);
2. Information that is expressed by automated data mining but is in a form that is not readily available for exploitation by automated data mining methods (e.g., a relationship between the ratio of two obscure customer aggregates that could be derived from a transaction stream and an outcome of interest in a customer table, such as a fraud tag);
3. Patterns that exist in the dataset being mined in a form that is accessible to automated mining, but are either incorrect, open to misinterpretation, or in some other way misleading (e.g., a strong but spurious correlation between a variable that is thought to be independent of an outcome of interest, but which is in fact causally dependent on the outcome);
4. Information that is not expressed in the data set being mined at all, but is either essential or highly relevant to producing and understanding meaningful patterns in the data (e.g., information about competitor activity, which is hard to capture in an analysis dataset, or knowledge that the basis of aggregation for a particular quantity in the dataset changed on some data during an observation window, giving rise to apparently changed behavior, when in fact number real behavioral change occurred).

This states that the consequences of the existence of these four kinds of information have direct and concrete implications for the construction of useful data mining tools, and that while most current tools concentrate strongly on information of the type that we listed above, most of the business value is likely to be lost if types 2–4 cannot be handled. This section states that

- Useful data mining tools must allow users to look at the raw data in powerful ways to allow the business analyst the opportunity to understand the data and spot deficiencies.
- Useful data mining tools must allow that business analyst to manipulate and preprocess the data in powerful ways to expose patterns that may otherwise be inaccessible to the mining methods available
- Useful data mining tools must express patterns found in ways that are comprehensible to the business analyst to reduce the likelihood of producing statistically sound patterns in the data that in reality are meaningless, false, or require careful interpretation.
- Useful data mining tools must allow the business analyst a variety of methods for controlling the mining process to allow information not present in the data either to be incorporated in the mining or to influence the mining in ways to compensate for the lack of information in the data.

- Useful data mining tools must express patterns in ways that maximize their action ability, if analysis is to be of practical business value, rather than just a source of "did you know...?" statements.

The above suggests that a useful data mining tool should aim, first and foremost, to empower the business analyst to explore and understand the dataset in relation to his/her own knowledge, rather than aiming to replace the analyst with some automated data discovery algorithm. Such a tool would necessarily provide an integrated set of functionality to facilitate the full range of activities necessary in a meaningful analysis, with an emphasis on interactivity, visualization, and flexibility. These characteristics are at least as important in determining the utility and power of data mining package as the selection of a suitable set of methods from the ever-growing plethora of automated classification, discrimination, and clustering algorithms.

Summary

We contend that data mining techniques should be evaluated according to the business task. This requires knowledge of the business process and interaction with the end users. Although most traditional evaluation has held time constant, the time variable cannot be forgotten when data mining software is put into the business process. The learned models can be evaluated and compared along the time dimension. By understanding the characteristics of the learned models, developers as well as end users can make more informed decisions.

The data mining process must be driven by those with expert knowledge of the domain. This has many implications for the process and for the tools, which support it: the process must be thoroughly domain oriented rather than technically oriented, and the tools must support an interactive, incremental, and iterative style of work. Data mining, because of its interactivity and domain orientation, has sometimes been described as a "hunch machine." The key to commercial success in data mining is *hunching, not crunching.*

Review Questions

1. Explain how data mining is used for business problems.
2. What are the benefits of a standard and common data mining process model?
3. Describe data mining and business knowledge strategy.
4. Explain the success of data mining in business with a case study.