

DATA MINING FOR PRIVATE TREASURE: SUCCESSFUL BUSINESS POLICIES INCORPORATE ETHICS

Dinah Payne

Department of Management, University of New Orleans
2000 Lakeshore Drive, New Orleans, LA 70148, USA
dmpayne@uno.edu 504-280-6961

Brett J. L. Landry

College of Business, University of Dallas
1845 East Northgate Drive, Irving, TX 75062
blandry@udallas.edu 972-636-8633

ABSTRACT

An increasingly popular business practice, data mining provides great societal benefit in many ways. It allows for the extraction of information from existing data to identify trends in consumer purchasing practices. It promotes knowledge discovery by helping companies to more fully understand the information it should be seeking to better serve the needs of consumers. Data mining fosters greater efficiency in companies' marketing efforts. There are corresponding costs associated with data mining, as well. The most difficult issue surrounding data mining is that of individual privacy rights and the costs associated with the potential alteration of "traditional" privacy rights. This paper seeks to review basic definitional information on data mining. Additionally, it seeks to provide a survey of arguments for and against data mining in general. Finally, a strategy for companies' successful, meaningful and ethical use of data mining is presented.

INTRODUCTION

Currently, society has two efficient and accurate sources to a wide spectrum of information: data mining and its natural extension, text mining. The information potentially available to anyone with the right resources ranges from people's name, address and telephone number to their personal financial and medical information (Montana, 2001). "Seeing" this data is becoming a major component in decision support and the formation of organizational strategies. The "Age of Information" has transformed society's understanding of the concept of information from its present form of "something that we ought to know" into its evolving form of "something we could otherwise never have known or used effectively."

Data Mining: Definitions and Function

Data mining "attempts to extract even more information from existing data by finding a correlation or trend in the (existing) data. It is also called knowledge discovery " because data miners do not know specifically what they are looking for before they find it. They are seeking to discover new insights from the data in their databases (Cary, Wen, and Mahatanankoon, 2003, p. 158)." Text mining is a more refined type of data mining, one that allows for more intelligent, refined and efficient searches of textual information. The focus of this paper is on the broader

search model of data mining. Data mining can be defined broadly as the analysis of database information. By its simplest definition, data mining is the set of activities used to find new, hidden or unexpected patterns in data (Marakas, 1999). Its purposes can be described in a variety of ways. First, data mining is the analysis of data that helps firms establish relationships and identify patterns of the subject of the data search (Modern Practice, 2005). Database information mining can also be used to identify specific product information and codes. Further, purposes of both cleansing and re-formatting data for future use are served by data mining (Supply and Demand Chain, 2005).

Further, data mining is used as an information extraction activity which has the goal of discovering hidden facts contained in databases. Using a combination of machine learning, statistical analysis, modeling techniques, and database technology, data mining finds patterns and subtle relationships in data and infers rules that allow the prediction of future results. Typical applications include market segmentation, customer profiling, fraud detection, evaluation of retail promotions, and credit risk analysis (Two Crows, 2005). According to Marakas (1999), data mining processes may be classified by the functions they perform or by the class of application they can be used in. Four categories emerge: classification, association, sequence, and cluster. For example, the response of customers to a particular direct mail campaign can be predicted by designating the appropriate parameters in the firm's information database from which the direct mailing campaign information will be drawn.

A second function of data mining is the association function. This approach employs linkage analysis of transactions that have a high probability of repetition. For example, the purchase of two products from the grocery store by the same person at the same time can be associated, such as skim milk and yogurt, peanut butter and jelly. The sequence function of data mining relates events in time. A very good example can be found in the mail catalogue industry: when a customer uses a credit card to buy clothes from a certain type of store, data mining allows other, similar companies to be apprised of the purchase. The firm who mined the data about the purchase will then immediately target that buyer. A rational extrapolation of the first purchase is that the consumer may be likely to make similar purchases from other merchants. This function of data mining appears to be the most important reason that businesses use data mining. It is used as a tool for identifying consumers to target in direct marketing campaigns (Cary et al., 2003). Finally, the fourth function of data mining is simply to identify sets of objects grouped together by virtue of their similarity or proximity to each other. This approach might be used to mine credit card purchase data to discover that meals charged on a business card are typically purchased on weekdays and have a value greater than \$250, whereas meals purchased using a personal credit card occurs on weekends and have a value less than \$175.

Common Usage and Stakeholders

There are many uses to which data mining can be put. Additionally, there are many stakeholders who both use and are affected by data mining. Data mining tools for these uses include a wide range of analytical activities, including data profiling, data warehousing, online analytical processing and enterprise analytical applications (Agosta, 2003). The most common uses of data mining efforts can be categorized into four areas: efficiency, security, customer-service and product innovation. It can be used to increase efficiency and enhance security via its ability to

detect fraud, waste and abuse; the discovery of patterns of monetary disbursement can reveal both inefficient spending patterns, as well as unauthorized patterns of spending. Additionally, efficiency is enhanced by the use of data mining via the better management of human resources; for example, employers may be able to choose better employees for specific jobs than they were able to without the knowledge base provided by data mining. Data mining can be used to improve service performance for customers; patterns of customer choice and spending can help the firm better provide more satisfactory goods and services for customers.

The final category of use of data mining is enhanced product innovation. Like the use of information to provide better customer service, data mining can provide better analysis of scientific and research data. For example, drug and medical research can be greatly enhanced by data mining techniques. Today, thanks to data mining technology, business is easily able to collect volumes of information, store it and access it at any time to mine for the necessary information. Also, this technology has enabled business to perform data mining on many types of data, including those in structured, textual, web or multimedia forms. In addition, the data mining techniques can be implemented rapidly on existing software and hardware platforms to enhance the value of existing information resources and can be integrated with new products and systems as they are bought on-line. When implemented on high performance client/server or parallel processing computers, data mining tools can analyze massive databases to deliver answers to questions such as "which consumers are most likely to be receptive to our advertisements?" or "which industrial processes are more likely to be most efficient?" (Cary et al., 2003).

In order to understand the benefits and burdens of data mining more fully, one must recognize those who are most closely affected by its use. In this review, it is imperative that the interests of stakeholders be considered; an ethical analysis of data mining must begin with the identification of the key stakeholders affected by data mining practices and a description of how the stakeholders may be positively or negatively affected. "A stakeholder is any individual, group, organization, or institution that can affect, as well as be affected by, an individual's, group's, organization's, or institution's policy or policies (Wood-Harper, et al. 1996, p.9)." These include, but are not limited to, customers/clients of the data warehouse, data warehouse management, subjects of information searches, shareholders of all companies affected by the firm's behavior, employees of the various firms involved, the community, society at large, and current and future financial backers of the firm. Other stakeholders include professional associations, governmental regulatory agencies, competitors, and information suppliers. In the consideration of data mining, several constituents are more important than others (Payne and Landry, 2005). The magnitude of harm that might be felt by the stakeholder if the data mining process was not executed in an ethical fashion could be great. For example, the public image of the firm collecting the data can be significantly harmed if consumers feel that the collection of the data was an egregious violation of their privacy. On the other hand, stakeholders such as professional associations monitoring the use of data mining or future financial backers of a project may not suffer the same degree of harm: in this case, the data miner might be less likely to be perceived as the "bad guy" in data mining efforts.

Business Benefits and Ethical Burdens of Data Mining

There are large numbers of uses to which business entities put data mining. These uses correspond to the many benefits accruing there from. Commercial information services have been in existence for decades, providing services regarding financial information (O'Harrow, 2004). Use of data mining in this area has arguably reduced the number of financial failures among unsuitable borrowers. Another benefit of data mining is that derived from better marketing practices. "Data mining and direct marketing are beneficial to the business community because they enable businesses to identify more accurately the target audience for their product or service, thereby reducing marketing costs (Morse and Morse, 2002, p. 77)." Corporations can use mathematical and statistical techniques to determine salient behavior patterns that were previously hidden in large databases compiled by the business (Markoff, 1999; Morse and Morse, 2002). According to Morse and Morse (2002), there are two major benefits that accrue as a result of data mining. First, the discovery of relevant data and behavior patterns allows marketers to better understand the interests and purchasing behavior of consumers. This benefit is the foundation of the second benefit: the more accurate knowledge marketers have about their consumers, the less money they will have to spend either in identifying the consumer or in identifying the consumer's consumption habits or patterns. "Businesses can save money on marketing while increasing their customer base substantially (Cary et al., 2003, p. 158)."

While business is using data mining for improving service, detecting fraud, analyzing scientific information, and, *inter alia*, managing human resources, it can be concluded that vast amounts of data, including personal information, can be collected, organized, and manipulated easily in the firm's efforts to uncover hidden consumption patterns and predict future trends and results. Thus, although there are many significant benefits to be derived from data mining, there are also serious drawbacks to unlimited or unregulated use of data mining. Incorrect conclusions can be drawn from data, data can be used for other than the original purposes for which it was collected and privacy rights can be violated. Additionally, data mining can create inferences that reveal information that the subject of the information does not want or choose to reveal: information that could be harmful if known to the wrong person. Further, costs of data mining must include those attributable to the improper or even incorrect collection and use of the data. "Data mining is not carried out with scientific rigor. The quality or randomness of the original data is not strictly verified and therefore the significance of inferences drawn from the data must be in question (Cary et al., 2003, p. 161)." The potential costs of data mining and text mining can be substantial and unexpected. These costs can be both tangible and intangible and result from consumer opposition to data mining practices that may be seen as unethical. For example, Lotus Corporation had to abandon a potentially lucrative product called Households when public outrage about privacy right violations caused their stock to fall. Households was designed to search for customers' profiles according to data provided by Equifax, a credit reporting agency. The public forced the company to scrap the product because they believed that their privacy rights were being violated (Anonymous, 1993). Concerns over privacy rights violation in general give rise to two more potentially costly problems of data mining. First, the subject of the information search does not have control over his own data, yet it is he that will suffer from the inaccurate or incorrect assessment of the data. Second, the gatherer of the information must give heed to his responsibility to keep the gathered information secure. "If companies are going to gather and infer sensitive data about individuals they must make a reasonable effort to protect it

from unauthorized access and from unethical use by employees or outside agents (Cary et al., 2003, p. 162).ö These problems are encompassed in the larger question of privacy rights of those about whom the information is gathered. The question is then begged: how can the technology of data mining be used without violating privacy rights?

The increased use of data mining raises many concerns regarding privacy. There is a growing concern among consumers that the right to privacy is being eroded by the increased sophistication of data collection and mining practices by both corporations and government entities. Exacerbating the consumersø concern is the question of ownership of the consumersø personal data. These ethical questions need to be identified and addressed by businesses and the government whenever they create new applications of data mining. This essay focuses on business rights and responsibilities surrounding data mining, rather than governmental uses of data mining. Business is confronted by many ethical questions in the use of data mining. These questions should be considered at all stages of the process, from when the original data is collected to when the insights gained from data mining are put to use. Companies today are gaining more of the consumerø personal information without the average consumer even being aware of the collection or transfer of this data. An efficient method of data collection is through the use of the Internet. Technology has made it possible to electronically monitor a personø interests, beliefs, purchasing habits, the kind of people they talk to and type of lifestyle they lead. One method of data collection uses web cookies to assign customer identification numbers. Additionally, at all purchase sites, including the tangible ones at the mall, electronic databases record and track purchases, consumer names, addresses, credit card information and so on. The use of registration forms is being used through the Internet as well in the search for information. The information provided by the consumer in the registration form is kept, stored, retrieved and used as deemed needful by the data warehouse or purchaser of the information from the data warehouse (Morse and Morse, 2002).

Ethics concerns arise through the use of these data collection methods. Depending on the method of data collection, the individual is not aware that he is subjecting himself to being monitored, he has not been asked for his consent to the search nor does he know where this information will go. Some companies collect information only for the government and some sell the information to other companies. To further intensify the problem, the individual is not given any choice about future uses of the data that he provides. Several ethical issues are inherent in the collection and mining of personal data: privacy, consent, ownership, and security. Many consumers feel that their privacy is violated when the provision of information is a requirement for purchasing and that information is utilized in ways in which they did not explicitly consent. The companies claim the information they are gathering is a public good gathered in a public sphere and that therefore privacy is not being violated (Cary et al. 2003). However, the question is whether the information derived from the data is private. Such information about the customer is not actually supplied by the customer (Wahstrom and Roddick, 2001). In addition, the information may have been required in order for the customer to make purchases and that practice alone raises privacy issues. Further, firms do not make a sufficient effort to inform the user of the current or future uses of his data. The user or customer is not provided the opportunity to provide informed consent. Consumers may not be aware that the company may combine the approved information with public information and prior information gathered to create a profile of the customer.

Another concern is in the type of data being collected. Some types of personal information are seen as being more sensitive than others (Cary et al., 2003; Wahlstrom and Roddick, 2001). Many consumers are unaware that credit history, financial information, employment history and possibly some medical information are routinely sold. Privacy rights have been linked to many other rights deemed to be essential to the development of a well balanced individual and society (Levine, 2003). These rights then must be given the appropriate level of respect.

One final ethical issue surrounding data mining is the security of the data collected and mined. If anyone is going to gather and infer sensitive data about individuals, they must make a reasonable effort to protect it from unauthorized access and from unethical use by employees or outside agents. The questions that arise here are: how well are the privacy laws written and how well is the law actually followed?

Ethical Data Mining Suggestions and Strategies

There are two categories of tools that can be used to assure the appropriate use of data mining: technological tools and managerial tools of business ethics. Technological tools include things like anonymity tools and security measures like file encryption that prevent or secure data from being used without proper consent. Three proposals to preserve privacy rights are reviewed here. The first proposal, Montana (2001), is more legalistic in nature, guiding firms in the development of policies designed to prevent legal or public image problems associated with privacy violations. The second and third approaches (Cary et al., 2003 and Raiborn and Payne, 1990) are more general in a managerial sense, providing guidelines which, if followed, should not only preserve the legal and public image integrity of the firm, but should also enhance the firm's ability to defend its actions ethically.

All three approaches to fairly and legally using data mining as a business tool can ultimately be described as having three main thrusts: customer orientation and adherence to sound ethical and legal principles. These thrusts are relatively straightforward. The customer orientation stresses the importance of keeping the consumer happy by preserving his legal and perceived privacy rights. The adherence to sound ethical and legal principles provide the firm the foundation to defend itself, again, both ethically and legally, in the event of some question of its use of data mining. Finally, drawing on the proposals to legally and ethically mine data, this paper presents a code of ethics that should be applicable in any situation, including those fraught with questions of privacy rights.

Privacy was a sensitive issue long before the advent of computers. Concerns have been magnified, however, by the existence and widespread use of large computer databases that make it easy to compile a dossier about an individual from many different data sources. Privacy issues are further exacerbated (by how easy it has become) for new data to be automatically collected and added to databases (Cranor, 1999, p. 29).

Providing a pragmatic approach to the privacy problems generated by the use of data mining techniques, John Montana (2001) suggests that the firm follows a strategy consisting of engaging

in five actions to avoid legal trouble or public image damage. First, particularly in the customer orientation, the firm considering using data mining should consider the expectations of the persons whose information they are collecting and/or using. Cavalier disregard for the consuming public's expectations is likely to lead to dissatisfaction, a growing refusal to be used in this manner and a backlash of public policy in the form of the development of new and more restrictive laws governing the collection and/or use of data. Montana's second suggestion is one of adherence to sound ethical principle. It is for the firm to develop a privacy policy that clearly and immediately explains to the user whose information is collected what the information will be used for and what it will not be used for. This suggestion incorporates advice about the collection of consent from the user to use the information: it should not be such a draconian consent or registration form that the reader is likely to give consent to a use he does not expect without even realizing it. The third suggestion put forth by Montana has attributes of adherence to both ethical and legal standards. A firm should not resort to the use of complicated and "legal latin" type language and policies about the collection and use of the data. Confusing the consumer with legalese designed to provide loopholes for the firm are again likely to lead to an unpleasant backlash of consumer anger. Finally, Montana's final two suggestions are grounded in adherence to sound legal principle. Montana's fourth suggestion is that the firm fully understand its responsibility to abide by the law as it relates to the nature of the information collected and/or used. For example, there are stringent laws in place protecting the privacy of medical and financial information; the firm must respect the person's privacy rights as legally protected privacy rights. Finally, it is suggested that firms collecting and/or using data maintain constant vigilance with regard to the law and public opinion. Legally and managerially, compliance with what is legislatively and societally mandated is a business necessity.

Cary et al. (2003) have developed a strategy containing ten practices that would aid in the legal and ethical development and use of data mining systems.

The power and sensitivity of public opinion dictates that corporations act to self-regulate their practices related to the handling and use of personal data. Following existing laws and regulations alone will not be enough to protect a corporation from the risk of damage from a negative public perception of their practices (Cary et al., 2003, p. 163).

The authors here clearly understand that compliance with the letter of the law alone is not sufficient to prevent consumer dissatisfaction at the least and legal action at the most. The first strategic category for the development of an ethical data mining system is that pertaining to the customer. The first suggestion in this category is that the firm considers the expectations of the customer when beginning a new project that requires data collection or use. It is immaterial whether the firm is actually complying with the law if the consumer feels that his privacy rights have been violated. The development of a customer-oriented privacy policy is a closely related suggestion: consumers believe that to divulge private information is their choice and, as such, their wishes with regard to privacy should be respected. The firm should alert the consumer as to the uses to which the information will be put so that the consumer can then make a better informed choice about disclosing the information. Finally, the customer driven principle requires that the firm give more control to the consumer over what happens with the information

collected. To achieve this, full disclosure and honesty is vital when gathering the information itself and when obtaining the consumer's consent to use the information subsequently.

Cary et al. (2003) also present suggestions that fall into the strategic category or principle of ethics. There are several of these suggestions. First, the spirit of the privacy policy must be followed, not just the letter of the policy or the letter of the law. "Regardless of the depth or breadth of a legal(istic) code, every immoral or illegal behavior cannot be proscribed. Thus, the spirit of the law is always broader than the letter of the law (Raiborn and Payne, 1990, p. 17)." Second, the quality of the source data should be checked: data mining of wrong or inaccurate data can cause serious harm to reputations that could lead to other types of harm. As a custodian of the information, the firm arguably has a fiduciary duty not to disseminate incorrect or false information. Additionally, a corporate code of conduct should be developed to establish appropriate standards for practices and treatment of consumers. For example, such a code can prevent potential harm from accruing in the first place from the dissemination of inaccurate information. Finally, the firm should perform an ethical audit of the uses to which its data is put. This audit can help identify any ethical or legal concerns that may arise when data is used in new or questionable ways. It can protect the firm from public outrage or legal action by providing proactive guidance to prevent problems from happening.

There are three legal principles projected by Cary et al. (2003). First, the firm should research and understand all laws that may pertain to its activities, especially the law concerning information that may be considered sensitive, like financial or medial information. Additionally, legal procedural matters should be on the firm's "radar screen:" there may be federal, state, local and even international law that impacts the legality of the use of certain information in data mining operations. A second suggestion with regard to the legal principle of data mining policy is the requirement that the firm stay current on new legal and public policy developments, as well as new attitudes towards the collection and use of data. Finally, access to the data warehouse is of paramount interest here. The third approach utilized here is that designed by Raiborn and Payne (1990); they designed a methodology for creating a corporate code of ethics that was comprehensive, clear and enforceable. Using the standards of behavior and the values they suggested, it is possible to create a workable code of ethics. It should also work to help establish a good data mining policy. The four values presented in their model are integrity, justice, competence, and utility. The value of justice requires that fairness and equity are incorporated into the decision-making process. The theoretical standard reflects the highest standard of ethical behavior: this is the spirit of morality. The second level of ethical behavior is the practical level; it is the acknowledgement that the highest level of ethical attainment may not be possible in the world we live in. It reflects the use of extreme diligence in ethical decision-making: the decision maker should be as ethical as possible in the circumstances. The level of currently attainable ethical behavior recognizes the society, through the idea of the public policy, has certain minimum requirements for morality; here, the decision-maker does not strive to achieve heights of moral behavior, he merely seeks to satisfy the basic societal moral standard of behavior. Finally, the basic standard of behavior is that of the basic legal standard of behavior; in this instance, the spirit of the law is not a part of the potential solution. "The ethical challenge for business to be temperate in its use of the Internet is based on getting business to see itself and its technology as social agents as well as economic ones; in short, an ethical approach to data

mining and direct marketing must consider the social impact of new technology while its fiscal capabilities are cultivated (Morse and Morse, 2002, p. 94).ö

Table 1: A Composite Strategy for the Legal and Ethical Use of Data Mining

Principle/ Orientation	Montana’s Suggestions	Cary et al.’s Suggestions	Raiborn and Payne’s Suggestions
Customer	Consider consumer expectation	<ul style="list-style-type: none"> • Consider consumer expectation • Develop a customer-oriented privacy policy • Give consumers more control over their information 	Competence at the theoretical or practical level
Ethics	Develop a privacy policy Follow internal practices honestly	<ul style="list-style-type: none"> • Follow the spirit of the privacy policy • Evaluate the quality of the source data • Develop a corporate code of conduct • Perform an ethics audit to identify new uses of data mining 	Integrity at the theoretical or practical level Utility at the theoretical or practical level
Legal	<ul style="list-style-type: none"> • Adhere strictly to the privacy law • Follow changes in law and public opinion 	Research and understand laws/legal procedures surrounding data mining	Justice at the theoretical or practical level

References Available Upon Request