1020 Brock Road South, Suite 2008
Pickering, ON L1W 3H2
Phone: (905) 837-0005
Fax: (905) 837-2199
www.boirefillergroup.com

By Richard Boire December 2004

A Practitioner's Viewpoint on Data Mining and Privacy

Given the tremendous publicity surrounding privacy today as well as the voluminous amount of legislation to both ensure and protect the privacy rights of Canadians, my discussion on this topic will not attempt to convey nor even purport to offer any legal advice or opinion in this area. Essentially, my discussion will revolve around the myriad number of issues that data miners face on a daily basis regarding this topic of privacy. At the same time, opinions in some areas will be offered from one practitioner's viewpoint.

In order to properly begin discussing this topic, one needs to understand the current privacy legislation that is already in existence. The Personal Information and Protection of Electronic Documents Act, commonly referred to as PIPEDA, which came into effect on January 1, 2004, legislates all aspects of privacy beyond just the marketing discipline. Yet, the focus of this article will be on the marketing component of PIPEDA and more specifically how privacy impacts the practices of data mining within the marketing arena.

Before even discussing the privacy impact of data mining it is important to note that the vast majority of marketers have enacted privacy business practices long before the PIPEDA legislation came into effect. Why? Simply put, it makes good business sense to keep the customer happy. Keeping the customer happy implies a number of things that we need to do as smart marketers and data miners. But, first and foremost is the notion that we respect any and all wishes concerning their privacy. The PIPEDA legislation outlines principles that, if carried out faithfully by an organization, will help to ensure that the privacy of the individual is being respected. There are 10 principles which all need to be addressed by each organization as part of their overall privacy policy. In this article and the subsequent one, I am going to tackle three of these principles which most directly impact the field of data mining:

- Identifying Purposes or Use of Information
- Consent
- Security or Safeguards

Identifying Purposes or Use of Information

From a data mining perspective, this is probably the most significant principle that both data miners and marketers need to understand. How will the information be used? For instance, some initial considerations are whether the information is going to be used for acquisition of new customers or to target existing customers for potential customer marketing programs. This is a key consideration because the data environment is very different under both scenarios. For instance, in the acquisition environment, data is typically very sparse. In many cases, the only information or data that is readily available is data at an aggregated geographic level which includes levels such as enumeration area, postal walk, and even postal code. In all cases involving the use of aggregate postal area level data, privacy is not an issue since the information or data at that level represents either an average or median value of all the individuals who reside in that postal area. For example, one record at the enumeration area level may contain an income field containing a value of $55,000. This implies that all individuals residing in that enumeration area would have the income value of $55,000 attached to their record. The loss of individuality through our approach to using information in this manner eliminates any potential privacy concerns. This is further supported by the fact that Stats Can will report only those records and variable values where there is a minimum required number of persons living in that postal area.

However, there are data vendor companies that collect individual consumer data which includes both behavioural and attitudinal type information as well as name and address. Typically, these companies market to consumers by inviting them to fill out a survey. The companies can then earn revenue by renting a specific list of names and addresses based on the information which they have collected. Secondly, they can earn additional revenue by selling large groups of names with all this information to a third party. The third party may actually conduct the analytics in terms of building a model. Once the model is applied and scored against the data vendor's database, the data vendor earns revenue based on the names selected by the model. In this scenario, prospects, whether selected through some basic information currently available or through modeling, are identified as the ideal prospects using individual-level data.

Given that individual data is being collected, the successful vendors will have strict clauses within the actual research survey pertaining to how the information will be used. The consumer, recognizing the implication of how the data might be used, can then choose to either complete or not complete the survey. It is assumed that completion of the survey represents implicit consent by the consumer regarding how this information might be used. However, in many cases, companies will actually have opt-in consent clauses written right into the survey.

The above example illustrates that the capture of data at an individual level requires customer consent. With postal or aggregate area level data being used to target prospects for acquisition, no consent is required as no individual data is being utilized. We will deal with the issue of consent in a subsequent article.

For existing customers, the power of individual-level data to produce very profitable solutions is a data miner's delight. From a data mining perspective, he or she is keenly aware of how this information is going to be used. However, the underlying issue isn't about the data miner but the customer and whether or not he or she knows how this information is going to be used? Most organizations handle this type of communication by stating that the company will use a given customer's information in order to help market better products and services to that individual. There is no talk of taking the customer's information and building powerful statistical models in order to target them better for certain products and services, nor should there be. In fact, one could argue that nothing should be communicated regarding how one mathematically or analytically utilizes the information. What is important is that the customer knows that this information might be used to target him or her for more relevant products and services, nothing more.

The mathematics and powerful techniques that represent the tools to target customers are irrelevant to the privacy issue? Why? It is because of the way these tools are actually utilized in practice. Let's take a look at an example.

In building a response model, we typically need thousands of customer records to build this model. Without getting into the deep mechanics of how to build this model, let's talk about how one individual customer record might be used since it is individual-level detail that is of obvious interest to privacy advocates. For ease of simplicity, I will use myself as an example.

In building this model, I would take my record and first strip off my name and address but retain my postal code. The postal code of my address would be kept in order to allow the data miner to append Stats Can type demographic information. I would keep the customer number or some kind of customer ID that would allow me to link my information to other behavioural type tables that may exist within the company. Through this ability to link to other files, I can potentially append a vast amount of behavioural and demographic information to my record. However, from the data miner's standpoint, it is linked to a customer ID and postal code and not to an individual name and address. Now consider what happens next when we begin to conduct sophisticated analytics in actually building the model.

In building the model, we take Richard Boire's record along with thousands of other records and apply statistical routines in order to build the response model solution. In building models, it is important to understand what the math or statistical technique is exactly doing to the individual record. Fundamentally, statistics is based on how information varies on a number of records and attempting to measure this variation. What this means is that my information is grouped with all the other records in the particular sample that is being used to develop the model. The statistics then measure how all the response information (1 for yes, 0 for no) within this sample varies along with all the other information within this sample. An optimized solution is arrived when the variation of response is best explained by the variation of specific non-response behavioural information.

The privacy question to ask at this point, though, is: Am I using Richard Boire's individual level information for other individual type decisions beyond what was outlined in the privacy communication to me. The answer is no. At this stage, I am using Richard Boire's information in an aggregate manner in order to develop a statistical solution that can potentially target Richard Boire and other individuals for more appropriate products and services. But this targeting can only occur once the developed model is applied to all records that might potentially be used in promoting a certain product or service. Without any application of the model, no individual is being targeted for specific products and services as a result of this model.

However, the common practice is that most developed models will be applied against a database of customer records for some marketing campaign. By applying a model to the customer database, each customer record including my own receives a response model score. Although the scores may be unique for each customer record, it is important to understand that the level of decision-making is at a group level. In other words, marketers will select groups of customers based on a range of model scores. Separate communication strategies may be developed around groups of these individuals based on the model score ranges. In no case does the marketer develop an individual communication strategy based on an individual model score.

Even in the other types of solutions such as CHAID and Cluster Analysis, the output is even more transparent in the sense that solutions are in effect actual groups such as segments or clusters of customers. Segments or clusters which are developed by the data miner can then be selected by marketers. The marketers can then craft unique communication strategies to each of the groups based on the demographic and behavioural composition of each segment.

The use of this customer information by the data miner does provide tremendous economic value to the organization without compromising the privacy rights of the consumer. From the data miner's perspective, it is easy to observe that data miners are not interested in individuals per se but in their information and how it can be used to optimize a solution for a given group of records. From the marketer's perspective, the marketer may want to provide unique communication to each customer or individual. But the reality is that the information provided by the data miner enhances the marketer's ability to both identify and communicate to the best groups of people and not individuals.

In the next article, I will continue the discussion on privacy with the focus on customer consent as well as reviewing safeguards to protect customer data.

Richard Boire is a Partner with the Boire Filler Group, a database marketing consulting company that specializes in developing and implementing data mining strategies. He can be contacted at (905) 837-0005 or via e-mail at RichB@BoireFillerGroup.com.