The Lighter Side of Data Mining

As I write this column, I realize that summer is practically over. Labor Day is next Monday, the kids are back at school and for many of us, our summer vacations are becoming a distant memory. Recognizing that work demands as well as personal demands seem to increase for all of us after Labor Day, I thought I'd take a stab at what I call the "lighter" side of data mining. There won't be any tables, charts, graphs or presentation of any new concepts and theories that require a higher degree of concentration. Simply put, I thought I would write about some of the experiences over my collective years in the industry that have provided humorous anecdotes and that are sometimes more engaging in a cocktail party or social atmosphere. But I do want to convey to the readers of this article that I am no Jay Leno. Therefore, I will keep my day job and submit to you the reader that future articles will focus on more probing issues concerning the current practices of data mining within business today. I am sounding an awful lot like a politician, heaven forbid. Anyways, here goes and enjoy.

As technology has improved, significant strides have been made in facilitating the naming of variables. For instance, in the not so distant past, it was quite common for many applications to have a restriction on the number of letters that could be used for a given variable. Programming applications could not run unless variables or fields abided by this limitation. In SAS, the most commonly used application, the restriction was 8 characters or letters. This posed a different kind of challenge as we attempted to describe a variable's meaning within 8 letters. For instance, in one technology company, we attempted to describe software products (SO) related to disk access (DA) and file sharing (SH) within the intranet infrastructure of the company (IN). The name, SODASHIN, to many of our internal analytics people sounded like one of Santa's reindeer (Prancer, Vixen, Blitzen, etc). In conducting any analytics for this client, it was not uncommon to hear that familiar phrase "Let's look at how that reindeer variable performed." The Yuletide season still brings back fond memories of Santa's 10th reindeer and its contribution to some of our analytical solutions.

Another good example of variable nomenclature was our attempt to describe certain StatsCan census variables. In one case, we were attempting to describe the % of persons (PC) that were employed in finance (F), administration (A) and real estate (RT). Putting this all together, the variable name became PCFART. As one would expect, it was a rather easy variable to remember in any analysis considering the multitude of humorous connotations that easily arose from such a

name. Of course, our humor was always constrained to the confines of our office. The last thing that we wanted to imply was that there was a "bad odor" to our work.

The current advancements in technology no longer require this creativity in variable nomenclature. The entire label description can be used as the variable name which can facilitate the analytical exercise especially from a presentation perspective. In a sense, we have sacrificed some humor for increased efficiency. Considering that most data miners, including myself, are not comedians, I'll take the latter.

The next examples are not direct experiences but rather scenarios or "war stories" that I have heard about in my travels. They certainly have a somewhat humorous vane to them when viewed from a historical perspective. Yet, at the time of their occurrence, senior executives were less than amused as the actual outcomes impacted their bottom lines in a negative manner. Both examples cited below again speak to the extreme importance of being able to implement solutions properly. In both examples, response models were built and then applied to target the most likely responders.

In the first example, the response model was developed in SAS but then scored and implemented in a different programming application. In this programming application, one of the nuances of the application was that negative model scores were placed at the top of the list from a rank ordering perspective. The scoring routine, in effect, placed the worst names from a response standpoint at the top of the list. Needless to say, the results were disastrous with a money-losing campaign due to extremely low response rates. However, the model worked in the sense that the model correctly rank-ordered names albeit in reverse fashion.

In the second example, the response model was developed in SAS and applied in SAS. The response model was developed using a logistic function. In scoring this type of function in SAS to generate the actual list for the campaign, the SAS programmer created code which was essentially incorrect. When applying a logistic function, the actual variable signs (-ve or +ve) which appear in the regression output before the coefficient of each variable needed to be reversed upon applying the solution as an exponential function. This was not done and once again, the top ranked names actually represented the worst names from a response likelihood standpoint. Similar to what was observed above, the model performed very well when looking at its ability to identify high vs. low responders. Yet the actual campaign incurred a very significant loss due to the incorrect placement of names on the list caused by the erroneous scoring of this list.

I can honestly state that both examples represented great learning experiences on the importance of establishing quality controls and checks within the implementation process. However, it was both painful and costly.

The last example comes from the area of placement within the data mining area. Back in the late eighties when predictive modelling for target marketing applications was still in its infancy, the New York American Express office put out an offer to hire "modelers." Obviously, they received some appropriate candidates in terms of their mathematical and programming capabilities. However, one female candidate who made it to the interview stage actually brought her

"portfolio" with her. This portfolio consisted of photographs of her experience as a "modeller." Needless to say, she was not hired but I would have loved to have been a fly on the wall during the interview session.

Well, that's all for now and maybe enough for most of you. On a more serious note, the next couple of articles are going to explore the ever popular topic of "privacy" as I want to focus specifically on its impact within the data mining area.

Richard Boire is a principal partner at the Boire Filler Group, a data mining consulting company.