Opinion piece

Do you own a red car? Unmasking unusual data predictors to gain insight

Received (in revised form): 26th July, 2007

Sam Koslowsky

is Vice President of modelling solutions for Harte-Hanks, Inc. (NYSE:HHS), a worldwide, direct and targeted marketing company that provides direct marketing services to a wide range of regional, national and international consumer and business-to-business marketers.

Journal of Database Marketing & Customer Strategy Management (2007) **14,** 266–270.
doi:10.1057/palgrave.dbm.3250063

Sam Koslowsky
Harte Hanks Inc.
55 Fifth Avenue,
14th Floor, New York,
10003-4301, USA
Tel: +1 212 520 3259;
e-mail: sam_koslowsky@harte-hanks.com;
Website: http://www.harte-hanks.com

It is difficult to turn the pages of most any data mining-related publication without the author admiring a new technique that purports to be the ultimate weapon for marketers. Indeed, some of these algorithms do provide reasonable though unremarkable gains. Others offer little that is new. Certainly, no one in the recent past has presented a quantum leap in improvement in solving data mining problems.

Well, if the techniques themselves cannot provide advancements, what then can a manager turn to for innovations in the analytics process? Before I begin discussing this important question, let me qualify that I myself am a user of several different data mining technologies. When an analyst chooses a specific tool or technique for an assignment, his or her choice is based on a number of considerations, none of which is discussed at length in this paper. Nonetheless, whichever tool is employed for the task at hand, there are opportunities to be innovative — and often that innovation starts with what's being analysed, the data themselves.

John Tukey, a renowned statistician, highlighted the 'important distinction between data exploration and data confirmation, emphasizing that many statistical methodologies place too great an emphasis on the latter' (http://en.wikipedia.org/wiki/John_Tukey). Thus, data exploration requires a practitioner to think creatively and leverage data in atypical ways

Tukey also stressed the importance of re-expressing two or more original variables to uncover new structure and relationships among data. These new dimensions could be interpretable and useful, as we'll see shortly in several examples drawn from the world of marketing.

It is this sort of creative intervention that distinguishes a run-of-the-mill analyst from more seasoned professionals. As mentioned, data tools most always are needed. But employing data imaginatively can make all the difference in uncovering unusual data relationships that enable true innovation when put to use. At a management science seminar, using data creatively was the subject of a luncheon meeting (New York Chapter, Institute of Management Sciences, 2003). To give a flavour of how some marketing scientists exploit data, let me summarise how one veteran statistician uses

AGE as a predictor among household data. In addition to the typical ways AGE can be used, the following additional out-of-the-box approaches were suggested:

- Adding up all ages of children to arrive at 'TOTALAGE'
- Adding up the parents' age to arrive at 'TOTALPARENTSAGE'
- Finding the differences in ages of children
- Finding the differences in ages of the parents
- Finding the age that parents first became parents

While some of these transformations may appear unusual, I think the reader can appreciate how a seemingly straightforward demographic item can be converted into potentially more useful intelligence, which, in this case, was deployed to help market an insurance type policy.

The innovative treatment of data can frequently make or break an analysis. An experienced modeller can jot down myriad ways to apply data when approaching data mining problems.

Let's examine three more recent applications of data that shows how resourceful data employment can make a big difference leading to original, even breakthrough, insights.

The cases have been redesigned in order to cloak the brands involved. The figures quoted are all, however, correct in terms of order of magnitude.

ADDRESS DATA

In recent years, managers have been using market research surveys to enhance their customer databases. It is not uncommon to ask recipients of the survey their particular attitudes or tendencies. Consider the high-technology firm that designed a questionnaire to gauge the interest of physicians for a new product. The survey measured the potential use of this service in the eyes of these medical practitioners. Approximately, 900 completed surveys were collated, and there appeared to be an interest expressed among a small, but solid, proportion of doctors.

The next question to answer was how to identify these individuals in the marketplace at large. After all, the survey pinpointed a group that would be interested. That is good. But the issue of finding others in the database that also would be attracted to this service is another matter.

There were five pieces of data that were available on both the survey responses and on the marketer's customer database. How to develop a usable model from these five available data elements was not an easy task. The hope was that some combination of these survey responses and existing physician data would provide some clues as to the eventual target for this product. These clues then could be applied to each customer record so as to flag the physician to his or her potential use of this new service.

For those involved in these sorts of 'multi-data' mining exercises, one would agree that they are among the most challenging. It's difficult enough to arrive at something meaningful with two or three data variables, but when these five data items include age, gender, address, years in practice and marital status, one can see the difficulty and complexity here.

Data on the address of each physician's practice appeared in four contiguous columns. They were labelled as address1, address2, address3 and address4. In practically all cases, the first field was fully occupied. Each successive address field had fewer and fewer filled-in address components. In an attempt to find a smoking gun — that is, some aspect of the address formatting that might appear to be a germane predictive element — we invoked a standard procedure that we hoped would discern something unusual for these numerous address fields. Sounds like a stretch, doesn't it? Imagine my surprise when we, however, discovered the following: Any address appearing in any of the four address fields that contained either the term 'SUITE' or 'FLOOR' appeared to be more associated with physicians that find this particular product appealing. The table below shows that 11.3 per cent of all survey respondents were interested in this product. Those whose practice address contained the key terms 'SUITE' or 'FLOOR', however, displayed greater potential. The index (20.3 divided by 11.2) of 181.25 suggests this could be a significant predictor. As the results of the model's deployment later proved, it was.

Per cent interested in product	Per cent with 'SUITE' or 'FLOOR' in address field	INDEX
11.20	20.30	181.25

As frequently occurs, finding a reason for such a peculiar data relationship became a matter of speculation. In this example, we believed that the appearance of 'FLOOR' and 'SUITE' may be associated with smaller physician practices, who happened to have a greater proclivity for this technology offering.

THE RED CAR

Like many marketers concerned about their customer's next purchase, automobile manufacturers seek to know when a consumer is likely to secure a new vehicle. Typical clues to solving this problem may focus on prior leasing patterns or a car's service history. While these items may be available to the modeller for predictive purposes, a seemingly innocuous data element also has proven to be beneficial. Who would believe that those with red vehicles actually kept their cars a longer period than others? Here too, hypotheses relating to the reason(s) for this stunning relationship were quite amusing. Most, however, surmised that the 'flashy' hue and the associated need to be seen publicly as a care-taking owner of the vehicle may very well explain this seemingly cause–effect relationship. The chart below highlights this unusual correlation.

Average time in years owner keeps his vehicle	Average time in years owner with RED car keeps his vehicle	INDEX
3.50	4.40	125.71

While I am in agreement that throwing in the kitchen sink into a data mining exercise is not a productive exercise — it's like saying the stock market rises and falls with the water depth of the Irawaddy River — I have found that for explanatory purposes, or for finding some unusual relationships, it is fascinating what an analyst can surface.

MISSING VALUES

Twenty years ago, there were essentially two or three methods that were most dominant for analysing marketing data, and dealing with missing data elements. Today, there are a plethora of textbooks and articles that offer many insightful approaches to tackle this constant concern. Here, we won't review these techniques but rather demonstrate how missing data itself proved to be — at least in one case — a vital predictive component that was validated in the analytics exercise.

The case involves an insurance marketer seeking to acquire new policies. Its primary means of acquisition involves renting prospect lists, modelling previous campaigns and selecting from these new lists names that meet the marketer's objectives. Data were plentiful. One hundred and twenty-six fields were available for this analysis, all of them sociodemographic in nature.

For each of these input fields, a missing data flag was constructed. So, for example, let's say the field names were A1, A2, A3 …. A126. Associated with each of these original data elements, we computed B1,B2,B3… B126. The value of B1 is equal to '1' when the corresponding value for A1 is missing. Otherwise, it is equal to '0.' B2 is calculated the same way. It gets a value of '1' when A2 is missing; otherwise, it gets populated with a '0.' This coding scheme continued for all 126 fields.

Here is an example of five fields with the original data.

Prospect ID Short name	Age A1	Gender A2	Rent/own A3	Income code A4	No. of children A5
1001	34	M	R	B	1
1002	27	MISSING	MISSING	C	MISSING
1003	MISSING	M	O	MISSING	0
1004	41	MISSING	MISSING	B	2

Four prospects are highlighted, with five data fields. A short name is assigned to each to these fields to clarify the illustration.

Now let's look at the 'new' missing data fields.

Prospect ID	Age	Gender	Rent/own	Income code	No. of children					
Short name	A1	A2	A3	A4	A5	B1	B2	B3	B4	B5
1001	34	M	R	B	1	0	0	0	0	0
1002	27	MISSING	MISSING	C	MISSING	0	1	1	0	1
1003	MISSING	M	O	MISSING	0	1	0	0	1	0
1004	41	MISSING	MISSING	B	2	0	1	1	0	0

The 'B' fields are either '0' or '1,' as described earlier. We next concatenate the B fields (listing the binary values as a single number). We get the following values.

B1	B2	B3	B4	B5	Concatenated
0	0	0	0	0	00000
0	1	1	0	1	01101
1	0	0	1	0	10010
0	1	1	0	0	01100

It is now the job of the analyst to determine whether the patterns that appear in the concatenated field are in any way related to response. While this may seem like a trivial problem with only five input fields, it gets somewhat more complex as we try to solve a problem with 126 input data elements. How do we determine the precise pattern(s) that may affect or predict response?

Enter text mining. While reviewing this technology is beyond the scope of my objective in this paper, a short introduction to this exciting capability is most certainly worthwhile.

Text mining seeks to uncover patterns, relationships and common themes from the plethora of data that are typically available in free-form text format. Data mining, on the other hand, uses more well-defined structured data formats to identify correlations and trends. Presently, text mining is benefitting from applications triggered by the reams of text data accessible to businesses. According to SPSS, a prominent provider of analytic tools and software, (http://www.spss.com/predictive_text_analytics/index.htm?source=homepage&hpzone=tech, 2007), 'an estimated 80 per cent of an organization's information is contained in text.'

A text mining application that is receiving an increased attention is in the area of customer service. With the variety of messages and communications passing a customer service agents's desktop, many firms are questioning whether these 'notes' can be intelligently analysed and leveraged to generate a pinpoint response that meets the customer's needs and expectations. These notes, typically captured in free-form structure, include much textual information. How can a customer service representative or, for that matter, trained business analysts summarise all this text into something useful? Text mining tools scour the documents and, for example, identify words, phrases or themes that may correlate with a typical or frequent sale or service query (to which an intelligent, immediate response can be generated), or may provide an early warning signal that a customer is about to churn.

A key approach to analysing free-form data is to employ what is referred to as information extraction. This technology locates critical expressions and relationships within text. Additional customer outcomes also may be appended in order to enrich the analysis. Hence, a subsequent customer defection may be correlated to certain phrases in an inbound communication.

Returning to our missing values example, by applying text mining and similar methodologies, we were able to extract nine distinct patterns among the millions possible that were related to response.

Our next issue was to hypothesise on why this correlation among the nine existed. As our original file was made up of many rented lists, we discovered that certain lists provided more of a lift, while others were less efficient. If this was the case, then all we needed to do to validate this finding was to look at the source field from the mailing label and match it to the list. Probing that field did not, however, provide the same insight. We investigated further. In doing so, we learned from the list provider that several of the provided lists actually were aggregates of several sources. So, we may have been using list 'X' with 850,000 names. But list 'X' was comprised of six separate sources. These separate sources were not known to us. When we examined the original list sources, we determined that only two of these six sources of aggregate list 'X' proved to be worthy.

Thus, discovering the patterns in the missing data proved to be a home run.

There will be no claims here that we've found a new data mining masterpiece, so don't go and throw out all that new data mining technology that's been accumulated over the years. They are important, and serve a vital function to managers and analysts, alike. On the other hand, do not believe that these tools, by themselves, provide all the solutions — they cannot. Rather it is the experienced analyst manipulating data in creative ways that offers the greatest probability of success. And the next time one sees an owner of a red car who maintains an office as a 'SUITE,' be sure he or she is on your mailing list.