KNOWLEDGE@WHARTON

MARKETING

Finding the Right Tool to Unlock the Power of Data

Aug 29, 2012

Global Focus, North America

In the last few decades, statisticians and computer scientists have produced a dazzling arsenal of extremely powerful tools to help managers translate data into business decisions.

Having access to a wide array of versatile solutions is not ordinarily considered a problem in the world of business. But the rise of "big data" has also brought along with it the explosion of mathematical models made possible by today's low-cost computing and storage platforms. Ironically, this poses a number of substantial challenges to managers trying to making sense of ever-growing quantities of information.

For example, says, Wharton PhD student Eric Schwartz, managers may be tempted to, as he put it, "flex their data-science muscles" and use a statistical model that is simply too complicated for the task at hand. The result, he notes, might well be that the model produces bad advice.

Alternately, managers may waste time trying to figure out which of several dozen possible models would be the most precise fit for the data set they are using. But the time it takes to play statistical guessing games after the analyses would be better spent running their businesses, Schwartz says.

"Wouldn't it be nice to be able to know, just from looking at the data, how complicated a tool you should use with it?" he asks. Such a "recipe," Schwartz adds, would have two benefits: It would allow managers to pick the "golden model", one that was neither too complicated nor too simple. And it would let them do so quickly, before having to undertake a lot of the more complicated analytical work.

That was the genesis of "Model Selection Using Database Characteristics: Classification Methods and an Application to the 'HMM and Its Children,'" which is currently under review in the premier academic marketing journal. (The "HMM" in the title stands for "Hidden Markov Model," a widely-used statistical modeling technique.) Schwartz's collaborators on the paper are Wharton marketing professors Eric Bradlow and Peter Fader, his dissertation advisors, who are also co-directors of the Wharton Customer Analytics Initiative.

"No one in the real business world has time to run a bunch of different models on a data set to see which one is best," Bradlow says. "We've come up with a way of picking the winner that is quite sophisticated in its science but quite simple in its practical application."

Some background for non-statisticians: A data set or a database can be anything from a retail chain's sales figures to the donor list from a charitable organization. Managers consult a data set when they need to make a decision, like whether a product should be discounted or whether a group of customers should be targeted with a special promotion.

The complexity of data sets has grown in parallel with progress in the development of tools to extract information from them. Lately, as computers have become more powerful, the number and sophistication of those modeling tools have skyrocketed. Some are simple functions built into every copy of Microsoft Excel. More complicated are the Hidden Markov Models of the paper's title, which require their own computer program.

For their paper, Schwartz, Bradlow and Fader took 64 different data sets that were representative of numerous real-world situations, from retailing, online media and elsewhere. With each one, they then ran four different models ranging in complexity from the simplest — a "Beta-Geometric Beta Bernoulli" model, which can be run in Excel and on an average laptop — to the most complicated, a full-blown Hidden Markov Model, which generally requires the use of a specialized programming language and takes much longer to run.

A typical data set might have two years' worth of information. For each of the 256 (4 X 64) variations of data sets and models, the computer was fed, for example, the first year of data, and then told to use the model to predict the numbers for the second year. The results were scored and sorted by accuracy. In all, the number-crunching required 24,000 hours of computing time if a single CPU was used. But because it was run in parallel across many different machines on Amazon's Elastic Computing Cloud (EC2), in connection with a research grant from the company, the job took just two days to complete. "Without Amazon's generosity, we would be a few years older by the time the analysis was completed," Bradlow says. "However, with the use of cloud computing, this shrinks down to two days. Amazon's investment in our work, the first major use of the cloud for large-scale marketing academic purposes, is not to be underestimated."

Hidden Simplicity

The different permutations of data sets and business models sound like something requiring a massive database just to keep track of. But the major surprise of the paper — indeed, the reason the authors believe they can help lighten the workload of every manager who works with data — is the great deal of simplicity beneath all of the apparent complexity.

Specifically, the data sets ended up clustering into a small number of different groups, each with easily identifiable characteristics. In one group — for example, talking about a retailer's sales — the database would be characterized by a steep decline in aggregate sales over time. This suggests a model in which customers might buy a product a few times, but then stop purchasing it altogether. Another group of databases can be characterized by a slight decline in aggregate sales without an extreme purchase concentration (e.g., less than 80% of sales come from the top 20% of customers). This suggests customers may keep switching back and forth between being frequent and enthusiastic purchasers and being less active buyers.

The good news was that each of those patterns can be linked with one of the four models tested by the researchers. In business life, it's relatively easy for managers to have an intuitive sense of which of the four main groups their data sets belongs to; but in this work, that information can be gleaned by looking at a simple chart. Armed with that knowledge, managers will now be able to use the results from the paper to pick with confidence which statistical model is the most appropriate.

“When people who work with data read our paper, I want them to think to themselves, ‘Wow, I didn’t realize that just by quickly summarizing my raw data, I could figure out which tool from my analytical toolbox was the right one,'” Schwartz notes.

The decision process involved in actually matching a particular data set with a particular model is described fully in the paper, and is easily accessible to someone with a basic background in statistics, Schwartz adds.

Picking the right model for a given data set has significant implications for business decisions. Schwartz says that choosing the wrong model can degrade accuracy of sales forecasts or behavioral targeting. The other bit of peace of mind delivered by the paper is that managers with relatively simple data sets can choose relatively simple modeling tools without worrying that they might be missing something in their analysis.

“There are some people who spend too much time worrying about complex models without thinking about the business value,” Bradlow notes. “And there are some people who don’t worry about it at all, but who should, because their fundamental business value depends on it.”