Data Mining Questions? Some Back-of-the-Envelope Answers

By Tim Graettinger

Data mining[1], the discovery and modeling of hidden patterns in large volumes of data, is becoming a mainstream technology. And yet, for many, the prospect of initiating a data mining (DM) project remains daunting. Chief among the concerns of those considering DM is, "How do I know if data mining is right for my organization?"

A meaningful response to this concern hinges on three underlying questions:

- Economics – Do you have a pressing business/economic need, a "pain" that needs to be addressed immediately?
- Data – Do you have, or can you acquire, sufficient data that are relevant to the business need?
- Performance – Do you need a DM solution to produce a moderate gain in business performance compared to current practice?

By the time you finish reading this article, you will be able to answer these questions for yourself on the back of an envelope. If all answers are yes, data mining is a good fit for your business need. Any no answers indicate areas to focus on before proceeding with DM.

In the following sections, we'll consider each of the above questions in the context of a sales and marketing case study. Since DM applies to a wide spectrum of industries, we will also generalize each of the solution principles.

To begin, suppose that Donna is the VP of Marketing for a trade organization. She is responsible for several trade shows and a large annual meeting. Attendance was good for many years, and she and her staff focused their efforts on creating an excellent meeting experience (program plus venue). Recently, however, there has been declining response to promotions, and a simultaneous decline in attendance. Is data mining right for Donna and her organization?

Economics

Begin with economics – Is there a pressing business need? Donna knows that meeting attendance was down 15% this year. If that trend continues for two more years, turnout will be only about 60% of its previous level (85% x 85% x 85%), and she knows that the annual meeting is not sustainable at that level. It is critical, then, to improve the attendance, but to do so profitably. ***Yes***, Donna has an economic need.

Generally speaking, data mining can address a wide variety of business "pains". If your company is experiencing rapid growth, DM can identify promising new retail locations or find more prospects for your online service. Conversely, if your organization is facing declining sales, DM can improve retention or identify your best existing customers for cross-selling and upselling. It is <u>not</u> advisable, however, to start a data mining effort without explicitly identifying a critical business need. Vast sums have been spent wastefully on mining data for "nuggets" of knowledge that have little or no value to the enterprise.

Data

Next, consider your data assets – Are sufficient, relevant data available? Donna has a spreadsheet that captures several years of meeting registrations (who attended). She also maintains a promotion history (who was sent a meeting invitation) in a simple database. So, information is available about the stimulus (sending invitations) and the response (did/did not attend). This data is clearly relevant to understanding and improving future attendance.

Donna's multi-year registration spreadsheet contains about 10,000 names. The promotion history database is even larger because many invitations are sent for each meeting, both to prior attendees and to prospects who have never attended. Sounds like plenty of data, but to be sure, it is useful to think about the factors that might be predictive of future attendance. Donna consults her intuitive knowledge of the meeting participants and lists four key factors:

- attended previously
- age
- size of company
- industry

To get a reasonable estimate for the amount of data required, we can use the following rule of thumb, developed from many years of experience:

$$\text{Number of records needed} \geq 60 \times 2^{\text{Number of Factors}}$$

Since Donna listed 4 key factors, the above formula estimates that she needs 960 records. Since she has more than 10,000, we conclude ***Yes***, Donna has relevant and sufficient data for DM.

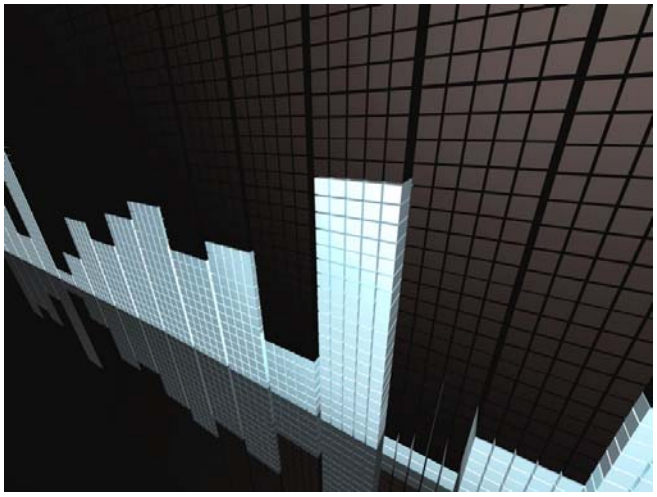
More generally, in considering your own situation, it is important to have data that represents:

- stimulus and response (what was done and what happened)
- positive and negative outcomes

Simply put, you need data on both what works and what doesn't.

Performance

Finally, performance – Is a moderate improvement required relative to current benchmarks? Donna would like to increase attendance back to its previous level without increasing her promotion costs. She determines that the response rate to promotions needs to increase from 2% to 2.5% to meet her goals. In data mining terms, a moderate improvement is generally in the range of 10% to 100%. Donna's need is in this interval, at 25%. For her, ***Yes***, a moderate performance increase is needed.



The performance question is typically the hardest one to address prior to starting a project. Performance is an outcome of the data mining effort, not a precursor to it. There are no guarantees, but we can use past experience as a guide. As noted for Donna above, incremental-to-moderate improvements are reasonable to expect with data mining. But don't expect DM to produce a miracle.

Conclusion

Summarizing, to determine if data mining fits your organization, you must consider:

- your business need
- your available data assets
- the performance improvement required

In the case study, Donna answered yes to each of the questions posed. She is well-positioned to proceed with a data mining project. You, too, can apply the same thought process before you spend a single dollar on DM. If you decide there is a fit, this preparation will serve you well in talking with your staff, vendors, and consultants who can help you move a data mining project forward.

Tim Graettinger, Ph.D., is the President of Discovery Corps, Inc. (http://www.discoverycorpsinc.com), a Pittsburgh-area company specializing in data mining, visualization, and predictive analytics. Contact Tim at (724)-743-3642 or tgraettinger@discoverycorpsinc.com

---

[1] For more background on data mining, see "Digging Up Dollars with Data Mining" at http://www.discoverycorpsinc.com