Grab Bag: Frequently-Asked Data Mining Questions and Answers

By Tim Graettinger

For the past year, I have presented a data mining "nuts and bolts" session during a monthly webinar[1]. My favorite part is the question-and-answer portion at the end. Participants with a diverse range of interests and experience toss out a lot of great questions. In this article, I'd like to share with you some of the best-of-the-best of those interactions. Some responses include pointers to other sources that I hope you'll find instructive and useful. Without further ado, let's get to the questions.

Question 1: What tools or tool sets do you recommend?

Let's start with an assessment – of your goals for data mining, of your team, and of your current tools and capabilities.

First, think about your goals for data mining. Are you planning a one-time project to solve a single business problem? Or, are you making data mining a corporate priority, and building a core capability within the enterprise? Your goals will have a profound impact on your choice of software tools and tool sets – obviously. On a related note, does your tool budget support your stated goals? (You do have a tool budget, right?) Be sure to include training and support costs in that budget. Allow me to make two more broad suggestions:

- Make a "wish list". Suppose that money is no object. What tools would you buy, and why? Really, write down your choices. This wish list is very useful. More on that later.
- Start slow. See what tools you actually need as you begin to do data mining projects. Acquire based on need rather than want.

Next, take a close look at the skills and background of the people on your team or team-to-be. Are they primarily MBA's or business experts? If so, you might want to consider industry- or application-specific software tools (e.g., for loan approval or fraud detection) that play to their strengths. Conversely, is your team composed

chiefly of statisticians and analysts? They might perform better with a more technical tool set – one that allows more control over data transformation and model selection (regression, decision tree, neural network, etc.). In any case, the key is to find the tools that fit the team's skills, that fit the way they work and think, and that fit the business and the planned applications of data mining and predictive analytics.

Third, consider the software tools you already have in-house and make an inventory. Compare the inventory to the "wish list" derived from your goals above. Look for areas of coverage, and then find the gaps. In my experience, data mining teams use a heterogeneous set of tools to do their work. I certainly use a variety of software tools[2], independently and together, to accomplish the following primary tasks:

- **ETL** (Extract, Transform, Load): to pivot, aggregate, and otherwise munge[3] source data files or database tables into a single, analytical file/table at the right unit of analysis for a problem.
- **EDA** (Exploratory Data Analysis): to compute statistics, create crosstabs, visualize, and ultimately interpret and understand the data.
- **Modeling**: to build regressions, decision trees, neural networks, and the like. Producing code to run the model in another environment (say, production) is an important consideration here.
- **Reporting and Presentation**: to present the results of a data mining effort and to monitor real-world model results over time. Useful tools might be as mundane as Excel or PowerPoint, or they might be very specific and tuned to your business operations and processes.

Finally, let's talk strategy for your tool search - now that you've thought about your goals, your team, and your existing capabilities/needs. I think a great starting point for your deep dive is KDNuggets.com. The site has software products and companies organized by industry and application, and they are identified as commercial or open-source.

I also urge you to join the discussion groups on professional, social network sites like LinkedIn and AnalyticBridge. You'll glean a wealth of insight from past postings. You can post your own, specific questions to persons with similar interests and similar experience in an industry or technology. Good luck!

Question 2: How do I get buy-in from management for Data Mining projects?

As usual, I have to speak from my own experience. So, please understand that my remarks are anecdotal, and not based on exhaustive studies of human or corporate behavior. That said, here's what I think and what has worked for me:

- **Find a "small pain"**: A pain is a business need or problem – one that you feel confident that can be addressed via data mining. I advocate that the problem you focus on be small – one that is relatively self-contained with a moderate risk/reward proposition. The problem should be one where you and your team are very familiar with the data source(s) and the business process. Go for one small success, then another, and another. Over time, it becomes easier to pitch a track record of modest successes than to make people forget a great, big, fat failure.

  Allow me to expound a bit more on this theme of a big, fat failure and what NOT to do. Too often in my career, I have parachuted into projects that were "in trouble". These projects shared some or all of the following attributes:

  - An inexperienced, newly-formed team over-reached and attempted to solve a very complicated, multi-faceted, high-risk problem.
  - The team did not focus on a business problem at all, but were lured into data mining by the notions of technical coolness and wizardry.
  - A team set about searching for "nuggets of knowledge[4]" within the vast mountain of data amassed by the enterprise.

  Look for these warning signs. You'll be glad you did.

- **Create a solid story line**: Craft a pitch that focuses on your business problem and how your data mining approach will eliminate the pain.
- **Be enthusiastic**: You are selling yourself and your team, not just the solution to a problem. Decision makers, the people who will fund your project (or not), need to be convinced that <u>you</u> can push your projects and programs to successful conclusions. Ask any venture capitalist, "What do you look for in a 20-minute pitch?" They'll tell you that they are looking at the person as much or more than the particular idea/solution/technology being presented.
- **Be persistent**: Sooner or later, someone will say "no". Learn what the objections are, and find ways to resolve them. Network in and out of your organization. Find others who can encourage, guide, and give you honest feedback.

I hope you see from the above that much of my approach to getting buy-in is not unique to data mining or even to technology. For me, it was, and continues to be, crucial to observe and learn from other successful champions within an organization - and outside it. Ask these people to mentor you. Ask them to evaluate your approach and your pitch. You'll gain much more than just buy-in on a particular project or technology.

Question 3: How should I transform non-numeric data?

Finally, a more technical question to wrap things up for this article. Let's focus on a particular type of non-numeric data – category data. Categories are extremely valuable for making predictions about customer preferences and future buying behavior. But they can be unruly and difficult to work with, especially for analysis and modeling methods that typically crave numbers. Suppose you have health insurance customer data that looks like the excerpt in Table 1. Each row represents a policy holder, and each column is a categorical attribute. The columns include:

REIN STATE	EFT_ FLAG	SEX	DEP_ CODE	PLAN_ CODE
	Y		F	5777
Y	Y	M	T	3231
	Y	F	F	5777
	N	M	S	1900
	Y		D	9888
	Y		K	2564
Y	Y	F	S	1900
…	…	…	…	…

Table 1 - Health insurance data

- REINSTATE –has the policy ever lapsed and been reinstated?
- EFT_FLAG – is electronic funds transfer being used?
- SEX – gender of the policy holder
- DEP_CODE – what type of dependents are included on the policy?
- PLAN_CODE – which coverage plan is in effect

To make an element like DEP_CODE, for example, amenable to data mining and modeling, I prefer to use a "1-of-N" transformation. Let's see how that works in this particular case.

First, determine the number of different categories that exist for DEP_CODE. There are five: {D,F,K,S,T}. Thus, we will create five new columns that indicate the presence or absence of each particular category in the DEP_CODE column. Sometimes, these new columns are called "indicator variables." The result is shown in Table 2. Works like a charm for building models and doing further analysis and mining.

DEP_ CODE	DC_D	DC_F	DC_K	DC_S	DC_T
F	0	1	0	0	0
T	0	0	0	0	1
F	0	1	0	0	0
S	0	0	0	1	0
D	1	0	0	0	0
K	0	0	1	0	0
S	0	0	0	1	0
…	…	…	…	…	…

Table 2 - Indicator variables for DEP_CODE

That's All the Space We Have…

In this article, we looked at several challenging, frequently-asked questions from my monthly webinar:

- What tools or tool sets do you recommend?
- How do I get buy-in from management for Data Mining projects?
- How should I transform non-numeric data?

I hope you found my responses helpful to you. In future installments, we will look at more questions from practitioners and business users alike. If you have a particular concern or question of your own, please feel free to get in touch – my contact information is below.

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.

Your comments and questions about this article are welcome. Please contact Tim at (724)-743-3642 or tgraettinger@discoverycorpsinc.com

---

[1] The webinar is Data Mining: Failure to Launch – and it's free. Sorry for the shameless self-promotion.
[2] If you want to know what specific tools I use, give me a call or drop me an email. I thought about naming names in this article, but decided against it.
[3] Manipulate sounds so boring compared to munge.
[4] My personal experience of pain and gnashing of teeth in one of these "nugget hunts" will one day become the subject of another article – if I can stand to fully recount the episode.