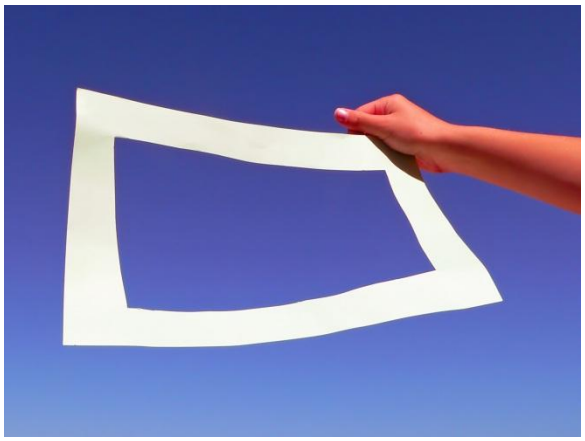
Framing the Data Mining Problem – Part 1

By Tim Graettinger

Where in the data mining process do humans like you and me – the data scientists[1] – add the most value? Is it in exploring the data to uncover anomalies and to fathom the relationships between elements? Is it in selecting transformations for the elements to improve their representations for modeling and analysis? Or, is it in building very sophisticated, nonlinear models that predict a future outcome given currently available data? Drum roll please while you consider your response …



No doubt all of the above are important and valuable. But for my money, the answer to the value question is none of the above. I believe, as data scientists, you and I contribute the greatest value by framing the problem. You can do a stellar job of exploring the data, transforming it, and building a model – but if the problem is framed poorly, it's all a pointless exercise. You're solving the wrong problem. Conversely, though, even a mediocre model - applied to the right, well-framed problem - will provide immediate benefit to your organization.

What do I mean by "framing the problem"? Simply put, it means to clearly, explicitly define what the problem is and is not. When I frame a problem, I work through a checklist that includes these five key questions:

- What is the unit of analysis?
- Who/what is the population of interest?
- What is the outcome?
- What is the time frame?
- How will we measure success?

By answering these questions, we frame the problem. We will look at the first two of these "framing questions" in depth in the rest of this article. We will tackle the latter questions in the next article in the series. Let's get started!

Question #1: What is the Unit of Analysis?

Are you familiar with this term, unit of analysis (UA)? It's social-science jargon meaning the major entity that you are analyzing and modeling - that is, the "who" or the "what". In very practical terms, the unit of analysis defines the record, the row, in a dataset.

Clearly, specifying the UA is fundamental when framing a data mining problem. However, I believe that the UA choice – and it is a choice - is also one of the most-overlooked aspects of any data mining project. Why? Perhaps because the UA choice seems obvious. For instance, Consider these data mining applications for a minute:

- Identify the best candidates for large donations to a non-profit
- Select the best location for a new health club
- Identify healthcare insurance fraud/abuse

What would you choose as the unit of analysis for each of the examples above? Think about it. I've got time. For non-profit fundraising, did you say "the donor"? For the health club siting problem, did you choose "the club"? For insurance fraud, did you select "the claim" as the UA? These were my initial choices when I worked on each of these applications. They seemed obvious. As we'll see below, other choices turned out to be more appropriate for the clients' needs.

First, consider again the fundraising problem. Some non-profit organizations believe that the decision to make a large donation is a household decision, not an individual one. For that reason, my client and I settled on the household as the UA. This UA choice impacted the way we aggregated the historical donations as well as how we appended third party demographic and behavioral data.

Next, ponder the health club site selection problem. In this application, the client had an existing network of about 100 clubs. That's a good business. But it's not good as a UA for data mining, since 100 clubs translates into only 100 records for modeling and analysis. Since we knew where the members lived, we decided to use the census block group[2] in the market area around each club as the UA. Each market area consisted of roughly 50 block groups. With the block group as UA, we suddenly had more than 5000 rows of data with which to build models – a much stronger position.

Finally, let's talk about healthcare insurance fraud. A lot of work has been done to try to identify fraud at the level of the individual claim. There are complex systems already in place at most insurers to accept or reject claims, or portions thereof, according to a myriad of byzantine rules. Initial discussions with the client ruled out that approach. As we talked more, we found what they really wanted and needed to do was find systematic behavior among physicians and other providers that was abusive, or in the extreme, fraudulent. The result: we selected the provider as the UA.

To summarize, the choice of unit of analysis (UA) is extremely significant. It defines the record, or row, for the data set – and thereby controls the number of records that are available for analysis and modeling. My suggestions:

- When framing a problem, don't assume the obvious. Look deeper.
- If low on data, ask "Can we go finer grained?" to break the data into meaningful and more fragments (and records), as we did for the health club siting problem.
- If data is plentiful, ask "Can we aggregate in an interesting, useful way?" like we did in rolling up the data to the provider level for the insurance fraud application.

Question #2: Who/What is the Population of Interest?

Above, the unit of analysis defined the row or record. Let's turn our attention now to the question of the population of interest – that is, the collection of rows that will form the dataset we will model and analyze.

Who or what is the population of interest? Is it your entire customer base? Or is it just those who have made over $100 in purchases? Is it all of the field service requests, or only those in the southern region? Just as the unit of analysis is a choice, so is selecting the population for your modeling effort. And, once again, what appears obvious on the surface may not be so obvious below it.

What makes choosing and getting to the proper population challenging? Here are just a few perplexing aspects that I have encountered in my career:

- The population/dataset that you are handed at the outset of a project is often a "sample of convenience". That is, it's the dataset that just

happens to be lying around. Typically, no thought has been given to whether or not this sample represents any particular population of interest.

- The dataset covers history from the "beginning of time". For instance, a customer purchase dataset might include transactions that span all 20 years of your company's existence. During that time, though, customer attitudes and behaviors have shifted significantly. Further, in this same period, your company has changed its branding three times and completely overhauled its product mix twice. External economic conditions have also swung wildly, particularly in the last five years.
- Critical population segments are NOT included in the dataset. This issue is particularly insidious since you can't simply look at the data and see who isn't there. You actually have to think outside the data. Gasp! For example, when modeling retention or renewal, your population must include both active accounts and lapsed/non-active accounts[3]. But these non-active accounts are often dropped from the operational database for efficiency's sake. So, sometimes you have to go back and get the non-active accounts from system backups. Ugly? Yes, but absolutely necessary.

What can you do to hurdle these obstacles and build a solid, useful population and dataset for modeling? In my experience, there are two components to success in this area - one is conceptual, and the other is procedural. Let's start with the conceptual aspect.

Conceptually, it's important to understand that the process of sculpting the population for your application is iterative. Once you've established this position with yourself, you also need to establish it with your client. With that expectation, your interactions with your client can be productive and insightful. Real quotes from my clients include such gems as these:

- "Oh, I didn't realize those people were in there. We don't operate in those states anymore. We should cut them out of the analysis."
- "That's an interesting segment. Maybe we should build a model specifically for them."

Having conversations like these with your client are good "relationship builders". You demonstrate your growing understanding of the business. Your client learns more about the data and about the data mining process. Those learnings will bear fruit for both of you in future projects and applications.

Now, let's consider the procedural portion required for success. What does the process look like? I think of it as a spiral, generally trending inward as you whittle away unwanted members of the population. Occasionally, the process will

spiral outward when you find you need to add missing members of the population, like the non-active accounts mentioned earlier[4].

From a practical standpoint, here are a few tips and techniques that I use when sculpting a population into shape:

- Spend some time thinking about the ideal population: the one you would like to have if you could choose it from scratch. With the ideal in mind, you can think about how to chisel your actual sample to look like your ideal population as much as possible.
- Filter your dataset to get the population you want, don't delete members. You may find those members useful later for a different model or analysis. And the code you write to filter those members serves to document your choices. You'll also need this code when you implement your model in production.
- Do segmentation/clustering in parallel with any predictive modeling you do. I build decision tree models constantly. I find them terrific for identifying important segments (groups) in the population. Sometimes they are segments that need to be removed. Other times, they are segments that could or should be modeled separately (see below).
- Build multiple, small models for different population segments rather than trying to build a one-size-fits-all model for a large, diverse population. For example, split a physician population by specialty (cardiologist, podiatrist, etc.) rather than trying to build a model that works for all physician types.

Wrap-Up

In this article, the first of a two-part series, we discussed "framing" a data mining problem – what that means and the value a human data scientist brings to the framing process. In particular, we considered two of the five key questions from my own framing checklist:

- What is the unit of analysis?
- Who/What is the population of interest?

We stressed the need to look beyond the obvious. And we noted that these framing questions imply choices that the data scientist and client need to make consciously.

The next article in the series will address the remaining portions of my framing checklist, namely:

- What is the outcome?
- What is the time frame?
- How will we measure success?

Until then, send me your questions and comments about data mining and predictive analytics. I look forward to hearing from you.

Tim Graettinger, Ph.D., is the President of Discovery Corps, Incorporated (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] I really like this term, data scientist. Data miner, as a self-reference, just never really clicked with me.

[2] The census block group is a geographical unit used by the US Census Bureau, and a block group typically consists of about 1500 people. With a little bit of effort based on the member's home address, each member can be placed into a block group.

[3] If you have only active accounts in the dataset, guess what happens? Your model would predict that everyone is active – not a real interesting result from a business standpoint. And you don't really need to build a model for that.

[4] I tried several times to draw a picture of this spiraling process. They were all disasters, graphically speaking. I'm sure you can imagine the concept better than I can draw it. Just be glad I didn't foist one of my attempts on you in the article.