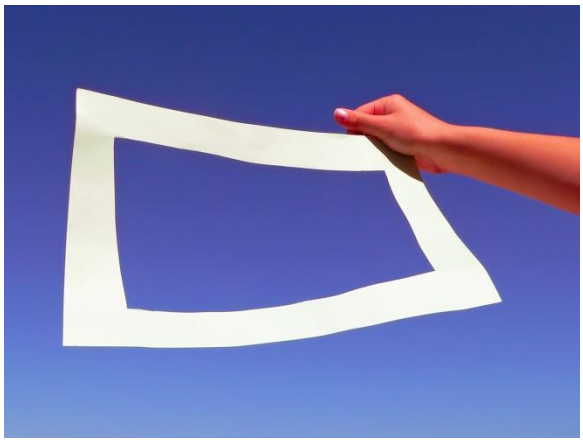
Framing the Data Mining Problem – Part 2

By Tim Graettinger

In Part 1 of this mini-series[1], we posed the question, "Where in the data mining process do humans like you and me – the data scientists – add the most value?" My response was that we contribute the greatest value by framing the problem well. In this context, framing means to clearly, explicitly, define what the problem is and is not. My framing checklist 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?



Part 1 addressed the first two framing questions above. In this article, we will tackle the remaining three topics. Keep in mind that framing is on an ongoing conversation between you and your client. It is a continual process of discovery and refinement. Nothing is more important to the success of your project, because…

The solution you build is determined by the way you frame the problem.

What is the Outcome?

Outcome. Output. Observed result. For us, these terms are synonymous. They all answer the question, "What happened?" For instance, in telecommunications, an outcome of frequent interest is renewal – did a customer renew their contract or did they terminate. In the world of non-profit fundraising, response to a marketing campaign is an outcome. That is, did a prospective donor make a contribution or not. Both of these outcomes have a yes/no flavor to them. Outcomes can be more diverse than yes or no, however. For a residential real estate application, we might choose the selling price, in dollars, as the outcome.

Perhaps the selected outcomes above seem fairly obvious, and in certain instances they are. Other times, the situation is not so clear cut, and a choice must be made – and the choice must be consciously made with buy-in from your client.

Consider the notion of renewal in fundraising. Gifts to a non-profit are freely made on a date chosen by the donor. There is no termination of services if no gift is made. Nevertheless, it is useful for organizations to think about donors who make gifts on a regular, "renewing" basis and those who lapse. Stop reading for a moment and ruminate on what outcome you might define for this situation.

You did stop and think for at least a minute, didn't you? Go ahead, you'll thank me later. In my work with various non-profit fundraisers, we have typically defined a "lapsed" donor as one who has not made a gift for at least 13 months. This choice is appropriate and pretty common in fundraising circles since many donors make contributions on an annual basis. Making the time window 13 months (rather than 12) gives ample room for the vagaries of human behavior.

Ponder a different application where now you are interested in finding new customers who look like your best customers. How might you define your best customers? Options abound, right? You might define best customers based on total purchasing dollars; say more than $250 in lifetime spending. Or, maybe the total number of purchases is a better metric for your business, and three or more purchases define a good customer for you.

These examples illustrate an important point about selecting and defining an outcome. It is very useful to parameterize the outcome (e.g., more than 13 months, greater than $250, etc.). This way, you can try various options by simply changing a number. By experimenting with the parameter value, you can find the one that makes the most sense and works best for your application.

A first note of guidance about choosing an outcome – be flexible. As mentioned above, be willing to experiment with your definition. And, be willing to make changes even late in the data mining process. Selling

price seems like the obvious outcome for the residential real estate application mentioned earlier. When I built a preliminary model, however, I found that the assessed value of the property was an extremely strong predictor – overwhelming almost all other factors. After discussion of this fact with my client, we decided it would be worthwhile to predict the percent difference between the selling price and the assessed value. The assessed value provides a "stake in the ground", a benchmark, for the selling price, and it also has a strong intuitive appeal for real estate agents. Our predictive model could then focus directly on other, finer-grained attributes that cause the selling price to deviate from the assessed value benchmark.

A second note of guidance – create a yes/no type of outcome, at least initially. For instance, try to predict whether a customer will purchase more than $250 in goods or services, rather than trying to predict the dollar amount of total lifetime spending. Why? First, building a yes/no classification model is simpler than building a continuous prediction model. Second, you will gain tremendous insight about the harder problem by solving the simpler one first. Finally, business processes often embed cut-offs anyway (e.g., sending special promotions to the customers with more than $250 in total spending). Why not design the cut-off in from the start, during the framing process.

What is the Time Frame?

In defining a lapsed donor in the section above, we alluded to a time window of 13 months. We also mentioned the notion of lifetime[2] spending when we defined our best customers. These are just two illustrations of how time plays a role in framing the data mining problem.

More generally, the dimension of time and the idea of time frames are fundamental to data mining and predictive analytics. In prediction problems, time enters the picture explicitly. You might be trying to predict the price of IBM stock two minutes from now, two hours from now, or two days from now. Or, you might be predicting the total dollars that a customer will spend over the next twelve months.

In yes/no classification problems, the time aspect is less obvious, more implicit. For instance, you might build a model to classify your customers into those who will renew their service contract and those who won't. No time component is called out explicitly. But, if you look closely, you will realize that this classification only has value for some time before a customer terminates or renews their contract[3]. Further, only information available before the outcome is known can be used to make the classification. It's subtle, but do these time-based distinctions make sense to you?

For you to frame a data mining problem, various time elements must be considered and reviewed with your client. These include the following:

- The time horizon for prediction. That is, how far into the future do you wish to predict the outcome? As mentioned above, do you want to predict stock prices two minutes, two hours, or two days into the future?
- The time window of relevant behavior. Here, we want to consider how far to "look back". Are the last 12 months of purchases sufficiently rich to predict the next 12 months? Or is it the last 6 months or 3 months that really matter.
- The time base of the population. In Part 1, we considered the importance of choosing the population of interest when framing the data mining problem. We can and should further refine our population by choosing a time base. That is, did long-gone customers from 10 years ago behave like current customers? Or, should you limit the time base to be customers who have joined in the last three years?

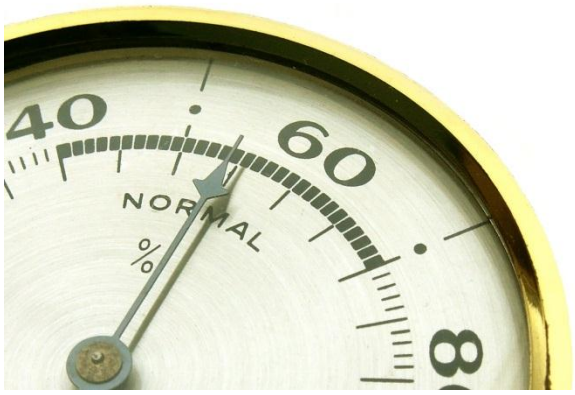
Some similar guidance applies here as it did above in defining outcomes. Once again, you will find it useful to parameterize - with numbers - your choices of time horizon, time window, and time base. In your software scripts and/or data flow and modeling diagrams, make it easy to adjust the parameter values[4]. In discussions with your client, call out these parameter values, and make her aware that they may evolve and change in the course of the project. Make it clear, though, that any parameter changes will be made only to better serve the business goals of the project – to make the right predictions at the right time.

How will you measure success?

Please don't say you are planning to measure success for your application using R-squared or Percent Correct. Please! These metrics are conveniently available right in the software tools, but they simply

measure model performance[5]. They don't speak to success in the context of the application.

What do we mean by success, then? That's exactly the question you need to answer when framing the problem. You need to define it – and parameterize it, too. Is it acquiring 10% more new customers than last year, or keeping 10% more of the ones you have? Is it 25% more profit or 15% more revenue from cross-selling and upselling your customer base? Is it finding at least half of the fraudulent transactions while keeping the false alarm rate below 1%?



Remember that the purpose of the predictive model is to improve the odds of achieving success. It is to rank the list of donor prospects so that the ones most likely to contribute are at the top. Or, it is to rank the list of insurance customers so that those most likely to terminate their policies are at the top. The purpose of the model is NOT simply to achieve 90% correct classification of the predicted outcome versus the observed outcome. Consider this: Suppose the termination rate is 10% per year for a health care insurance policy. A model that says no one will terminate their policy can be correct 90% of the time simply by predicting that no one will ever terminate. A model like that won't help your business achieve its goal of reducing customer churn, yet people get irrationally exuberant or irrationally disappointed about a percent correct number.

Defining success is the first component of this framing task. The second component is figuring out how to measure it. Why is measurement challenging? I think it's because the predictive model is always embedded[6] in a larger business process (e.g., renewal marketing), and it can be hard to distill out the impact of the model. Or, even more challenging, some models require a new business process to be developed because one did not even exist before.

For the former scenario where the model is embedded in a larger process, you might frame in a "traditional-versus-model" test and measurement strategy. That is, two groups of customers are selected to get the same treatment – the standard renewal marketing approach. The first group is selected "traditionally", say based on tenure. The second group is selected based on the rankings from the model. At the end of the renewal period, you measure the renewal rates for both groups.

If a new business process needs to be developed to measure impact and success, I strongly advise you to get help. Data scientists are not business process experts - at least I'm not (ask me how I know). Partner with an expert to define the process, to get buy-in from all the stakeholders, and then implement it when you are ready for roll-out.

Wrap-Up

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

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

With these framing questions in hand, you should be able to have productive and insightful conversations with your client - before you begin any data mining/predictive analytics project.

If you have any additional thoughts or questions about framing a data mining problem, call me or send me an email. My contact information is below. I hope to hear from you.

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] See "Framing the Data Mining Problem – Part 1", by Tim Graettinger.
[2] Where lifetime might mean the customer's tenure to date, or the last three years, or some similar type of duration.
[3] Once the outcome (renew or terminate) is known, you really don't need a model prediction of what is likely to happen. You already know what DID happen.
[4] Preferably, and make that a strong preference, design your scripts and data flow/modeling diagrams such that you only need to change parameter values in ONE PLACE.
[5] And they're not even particularly good at that. All that they are is convenient.
[6] The cheese does not stand alone. Ever.