Answers to Data Mining Questions, Part 1

By Jim Wheaton
Principal, Wheaton Group

Original version of an article that appeared in the November 1, 2004 issue of "DM News"

Many favorable emails were received as a result of the article, "Answers to Four Common List Questions" (*DM News*, November 4, 2004). Therefore, it was decided to repeat the format with a focus on a hot topic: how to integrate statistics-based predictive models into a coordinated, multi-channel contact strategy. This is the first installment of a two-part series.

Question #1:

We sell four services across two primary channels: direct mail and telemarketing. Also, we employ email as a customer-side support channel. We are developing a coordinated, multi-channel contact strategy across our four services. Therefore, we are replacing our outdated RFM Cells with regression models. For each promotion to a prospect, inquirer or customer, please outline how the models can help us answer: a) which service to offer, b) which channel to employ, and c) what the optimal channel mix should be, as well as the best timing and frequency across promotions.

Answer:

It is good that you recognize the general incompatibility of RFM Cells with modern database marketing. For companies with several services and multiple channels, the large number of cells that are generated by a typical RFM approach will result in what is know as a "proliferation quandary." You will end up with a choice: either too many cells to be practical, or too few to be effective. For a discussion of "cell proliferation" and the corresponding quandary, please see "RFM Cells: The 'Kudzu' of Segmentation" (*DM News*, July 15, 1996).

We will begin by noting that our answer is also appropriate for companies that offer multiple products. For example, it applies just as well to catalogers with several titles as it does to banks with multiple services such as home equity loans.

The answers to "a" and "b" – which service to offer, and which channel to employ – are based on the construction of a statistics-based model (or models) for each permutation of service and channel. The ultimate goal is to accurately estimate the profitability of each service/channel permutation.

One complication is that model scores are not directly comparable. The reason has to do with statistical theory that is beyond the scope of this article. For example, assume that a household's score for the Product X model is better than that for the Product Y model. Many non-data miners

will be surprised that, from a purely statistical perspective, this does not necessarily mean that the household should receive a promotion for Product X.

Fortunately, there are valid ways to compare models. A preferred method at Wheaton Group is to focus on financial projections tied to each model segment. This avoids any technical "landmines," and provides the added bonus of being a business-oriented solution.

The answer to "c" – what the optimal channel mix should be, as well as the best timing and frequency across promotions – is more a function of testing rather than predictive modeling. A series of well-constructed longitudinal test panels must be created, in order to calculate key metrics such as: 1) the amount of cannibalization across products/services, 2) the amount rate of cannibalization within products/ services by re-mails, and 3) the effect of time between promotions on these different cannibalization effects.

The mechanics for executing everything described within this answer – in an environment of multiple services (or products), channels and seasons – is far from trivial. However, it is an absolute requirement for arriving at contact management strategies that are data-driven and financially-focused.

Question #2:

How do I know when it is time to rebuild a model?

Answer:

It is likely that a model will have to be rebuilt whenever one of two things occurs. First, there has been a change in the underlying structure of the source data. Or second, there has been a change in the fundamental dynamics of the business; that is, when a totally different type of customer is being attracted to the product or service being offered. Models extrapolate from the past to the future, based on an assumption of environmental constancy. When there is a disruption in constancy, extrapolations become problematic.

Models generally are remarkably resistant to non-dramatic changes in creative and price. Therefore, as long as the fundamentals of the business remain reasonably stable, and there is no change in the structure of the source data, models are likely to retain their potency for years.

Fortunately, there is a way to determine the likelihood that model performance will deteriorate. Every time a model is scored in a production environment, profiles should be run on each segment. These profiles should include averages and – optionally – distributions for every one of the model's predictor variables. They should also include whatever RFM and/or demographic elements are helpful for "painting a picture" of the best customers versus the worst, as well as those in between.

These profiles should not diverge significantly from: a) profiles run off previous successful production mailings, and b) profiles run off the original dataset used to validate the model. The extent to which divergence has occurred is the extent to which model deterioration is likely to be encountered. Sudden, dramatic divergence generally is the result of a change in the structure of the source data. Gradual divergence often is symptomatic of a change in the dynamics of the business.