Building Your Model: A Primer for Choosing Variables

Last time, we discussed a variety of different types of predictive models for data. This time, we want to talk about what factors to consider when building a model and how to determine which variables are the most effective when creating a targeted marketing piece.

First, we must ask the question: Why evaluate data according to a statistical model in the first place? Simple, the more you know, the more targeted and effective your marketing program can be. But in some cases, the use of postal code indexes can be much simpler and easier to implement than in building a model. Let's take a look at an example.

If you know that you have 1,000 people in a certain postal code, and 100 of those are customers, your penetration is 10%. Therefore, if you are going to go after new customers, you have two options: 1) You can start where the penetration is the lowest, or 2) you can start where you've already been successful — in those postal codes with the highest customer penetration. Some people call this "fishing where the fish are."

Whichever option you choose, knowing your postal code penetration allows you to make an educated decision based on the data.

Choose Wisely

Presuming that we decide to develop a model, the issue of variable selection is critical to the process. After all, when it comes to sending out marketing pieces to customers, not all variables are created equal. For instance, a customer's geographic location might have some bearing on his or her behavior, but gender, age, or some other variable might not. In another case, a customer's geographic location might have no bearing whatsoever.

When creating a model, how do you know which variables to use and which not to? Techniques such as correlation analysis can be used to assess the impact of a given variable against the predicted behaviour as well as prioritize the importance of variables against this given predicted behavior.

Exploratory Data Analysis

Exploratory data analysis (EDA) is another way of predicting how variables will behave. Like correlation analysis, these reports demonstrate the impact of a certain variable against the predicted behavior. Listed below are two examples of EDA analysis, with one report leading to a useful variable and another where the variable was of no use:

Model #1	
Age	**Response Rate (%)**
< 25	0.50
26 – 35	1.00
36 – 45	1.50
46 – 55	2.00
> 55	2.50

Model #2	
Tenure (Years)	**Response Rate (%)**
< 1	0.75
1 – 2	1.25
3 – 4	1.00
4 – 6	0.90
> 7	1.05

From the above reports, we can see that older persons are more likely to respond to this particular campaign. This would indicate that age is an important variable and might likely be in a model. Conversely, when we look at the data by number of years as a customer, no discernable trend emerges. As a result, we would not expect this variable to be in the model.

Once correlation routines as well as EDA reports are run, the results lead to those variables which should be included in the more subsequent and more robust statistical routines such as CHAID, logistic regression, and multiple linear regression.

All of these methods are statistically valid. Determining which one is "best" for your campaign depends on how the model performs in that particular situation. Sorting a validation sample by descending model score, you can bucket the validation sample into 10% increments ranging from 0-10% (highest score names) to 90- 100% (lowest score names). We then observe the performance of each interval (decile) where the top-scoring deciles should exhibit better performance than the lower-scoring deciles. In other words, we are assessing the model based on its capability of rank-ordering performance from the top decile to the lower decile.

There are lots of ways to analyze your data and build models so that you are making the best of the information you have. There is no sense in "blind" marketing. This provides just one small example of analysis as it pertains to model-building that can be done today. For more information on using statistical analysis to refine the variables used in your marketing models, you should discuss this with a data mining professional.

Richard Boire is a principal partner at the Boire Filler Group, a data mining consulting company.