ASSESSING THE PERFORMANCE OF DIRECT MARKETING SCORING MODELS

Edward C. Malthouse

EDWARD C. MALTHOUSE is Assistant Professor in the Integrated Marketing Communications program, the Medill School of Journalism, Northwestern University, Chicago, Illinois.

ABSTRACT

Direct marketers commonly assess their scoring models with a *single-split, gains chart method*: They split the available data into "training" and "test" sets, estimate their models on the training set, apply them to the test set, and generate gains charts. They use the results to compare models (which model should be used), assess overfitting, and estimate how well the mailing will do. It is well known that the results from this approach are highly dependent on the particular split of the data used, due to sampling variation across splits. This paper examines the single-split method. Does the sampling variation across splits affect one's ability to distinguish between superior and inferior models? How can one estimate the overall performance of a mailing accurately? I consider two ways of reducing the variation across splits: Winsorization and stratified sampling. The paper gives an empirical study of these questions and variance-reduction methods using the DMEF data sets.

JOURNAL OF INTERACTIVE MARKETING
VOLUME 15 / NUMBER 1 / WINTER 2001

1. INTRODUCTION

Direct marketing scoring models are used to predict the future behavior of a group of customers. Consider an example. Suppose that a catalog company plans to circulate a back-to-school catalog to its customers and that it must decide which of its customers should receive the book. Sending a catalog to someone who is not interested in purchasing from the book is usually not profitable; therefore the catalog company would like to know who is likely to make a purchase. Scoring models can help the catalog company with this task, as well as many related tasks such as determining which prospects should receive a book.

Scoring models are usually built using historical data. In the back-to-school example, the company probably circulated a similar book during the previous year and observed who responded and who didn't. The company could use data from the previous year to make decisions about this year's mailing. It could use a predictive modeling technique such as regression to estimate how much each customer spent during the previous year, provided that the customer received the offer last year. Next, it would predict this quantity using purchase history it had on the customers prior to the mailing, starting with versions of recency, frequency, and monetary value. After estimating the model, it would apply the model to the current purchase history, called *scoring the database,* and have a better idea of who will respond to this year's offer. The functional form and estimation of scoring models has been the focus of much recent research (see, e.g., Bult, 1993; Bult & Wansbeek, 1995; Colombo & Jiang, 1999; Zahavi & Levin, 1997; Magidson, 1988; Hansotia & Wang, 1997; Malthouse, 1999).

This paper evaluates an important question that direct marketers fitting scoring models face: how can I assess the *performance* of a model? There are two reasons to ask this question:

- *Model selection: the relative question.* When a company builds a scoring model it usually ends up with several possible models and must choose one for implementation. To do this, it must know how one model performs *relative* to another. For example, the company may have used stepwise regression to select a subset of predictor variables for the final model; after using stepwise regression it must choose one of the resulting models. Alternatively, it may have tried different modeling techniques. In addition to using a regression model, perhaps it tried CHAID and neural network models as well; which is better? Also, the company could be evaluating whether or not to use overlaid variables, which it must pay to use; for example, the company could use zip-level Census demographics for free, or could buy more accurate demographic information. In deciding whether or not to purchase the more accurate demographics, it must evaluate whether or not they improve the performance of its models relative to those using zip-level data only.
- *The absolute question.* A second reason to evaluate a model is to estimate the performance of a mailing for planning purposes. How much demand will a particular circulation plan generate? In this case the objective is to understand how the model performs in *absolute* terms; for model selection the emphasis is on assessing the performance of one model *relative* to another. For example, the business plan for a company might specify that a certain number of customers must be "active" at the end of a time period. The gains chart for a scoring model will help predict how many customers will activate. Another example is assessing how a model will do on the margin. If one more book is mailed, what is the chance that this customer will activate. Such information is important in planning circulation across different campaigns, e.g., new customer acquisition, current customer retention, and former customer re-activation. In both examples there is a need to know how a model will perform in absolute terms.

This distinction is important because some of the methods discussed below will help modelers decide between models, but will give biased