

Superiority of Tree Analysis Over RFM: How It Enhances Regression

*By Jim Wheaton
Principal, Wheaton Group*

Original Version of an article that appeared in the August 12, 1996 issue of “DM News”

Last month’s “RFM Cells: The ‘Kudzu’ of Segmentation” showed why statistics-based predictive models – regression, neural networks, genetic algorithms, and the like – are superior to RFM Cells. The reason is that predictive models are more stable than RFM Cells, easier to implement, and substantially more powerful.

If RFM is inferior, then why is it so popular? One reason is fear of the unknown. Many of us are suspicious of statistics-based predictive modeling techniques because we don’t understand them. If we’re having difficulty understanding exactly what “Regression Deciles 8 through 10” are, for example, it’s going to be tough explaining to the CEO why we shouldn’t mail them!

For all of its shortcomings, RFM is understandable. Every marketer, for example, can explain why a customer with just one previous order – five years ago, for \$5 – should not be contacted.

But even database marketers who insist on an easy-to-understand, Cell-driven segmentation strategy should dump RFM. This is because there exist statistics-based tree analysis tools that are far superior.

One such product is “CHAID” (Chi-Square Automatic Interaction Detection). Another is “CART” (Classification and Regression Trees). This is not meant to be a product plug – I’m sure that there are other excellent tools in the market place with which I’m not familiar. P.C. versions of these products are available, and can be executed by marketers without an advanced degree in statistics.

Tree analysis is similar to RFM in that each node (“group”) is defined by easy-to-interpret “and” statements. Assume that we want to segment a customer file by response to a previous mailing. To keep matters simple, we’ll build a tree model off just six variable types: Months Since Last Order, Number of Orders, Average Order Size, Merchandise Categories, Age, and Gender.

The modeling process for tree analysis is a series of formal statistical procedures that divides our customer file into multiple groups – with response rates that are significantly different from each other. Each of these groups is defined by a subset of our six variable types. These subsets can be very different from each other. We’ll end up with – say – 31 groups, including the following (with their corresponding performance):

- 40-50 year old female jewelry buyers, with four or more purchases, averaging \$500⁺, and at least one purchase within the past six months – 8% response rate.

- 30-35 year old male electronics buyers, with three or more purchases, and at least one within the past twelve months – 4% response rate.
- 18-25 year old male sporting goods buyers, with just one order, 38+ months ago, for under \$15 – 0.5% response rate.

Clearly, these groups are no more difficult to understand than RFM Cells. They'll be just as easy to implement. But, they'll be much more stable than RFM Cells because they're the product of a formal statistical process.

Having said all of this, our research group prefers regression rather than tree analysis as its primary tool for predictive modeling. Regression has its advantages when the goal is to rank-order a prospect or customer file on predicted future behavior, not the least of which is ease of implementation.

Nevertheless, we are heavy users of tree analysis as an enhancement to regression. For one, tree analysis is a wonderful tool for creating what are known as interaction predictor variables. These are defined as the synergy in explanatory power that often results when multiple variables are combined. Frequently, such synergistic interaction variables are important components in subsequent regression models. The following example explains how this works:

- Assume that we're selling Florida beachfront condominiums, and that we're building a regression model to rank-order a prospect file by likelihood of purchase. While performing the up-front exploratory analysis that's required for any successful model, we notice that older individuals have a response rate that's moderately higher than average. This makes sense because retirees often settle in Florida.
- Then, we discover that affluent individuals also have a moderately higher response rate. That's not surprising either. After all, beachfront property isn't cheap.
- In a preliminary step before the actual regression, let's assume that age and income are entered into a tree analysis – along with many other potential predictor variables. As described in the previous example, the tree analysis will divide our prospect file into several groups – with response rates that are significantly different from each other.
- Let's say that one of these groups generated by the tree analysis is 65-72 year olds with incomes exceeding \$110,000. And, that this older, affluent group has a response rate that's seven times higher than average. This, by definition, is synergy. Two moderately predictive variables have been combined in such a way that an extremely responsive niche within the prospect universe has been discovered. This is what statistics-based predictive modeling is all about!

Tree analysis can also enhance regression models by providing insight on how to tailor the promotional message to the characteristics of a given prospect or customer. In this approach, the regression model determines whom to promote. Then, the tree analysis provides insight into what

the promotional message should be. This two-step segmentation strategy is central to many state-of-the-art database marketing programs. We'll discuss how next month.

Jim Wheaton is a Principal at Wheaton Group, which specializes in direct marketing consulting and data mining, data quality assessment and assurance, and the delivery of cost-effective data warehouses and marts. He is also co-founder of Data University. Jim can be reached at 919-969-8859, or jim.wheaton@wheatongroup.com.