New Statistical Tools for Key Driver Analysis

By John Colias, Ph.D.

Key driver analysis is used by businesses to understand which brand, product, or service components or attributes have the greatest influence on the customer's purchase decision or a physician's prescribing decision. The analysis can be based on statistical measures of the relationship between each attribute and an overall measure of the market performance of the brand, product, or service.

For example, consumers might rate a durable good based on quality, durability, and reliability. Physicians might rate a drug based on efficacy, side-effect profile, dosing, and patient compliance. Overall measures of market performance might include likelihood to recommend (for a durable good) or actual percent of prescriptions written (for a pharmaceutical product).

The purpose here is to discuss the potential application of a relatively new tool, Ensemble Prediction, which combines thousands of regression models to produce a prediction of the overall market performance based on attributes which influence the purchase decision or physician's prescribing decision. As it relates to key driver analysis in marketing research studies, Ensemble Prediction delivers an extremely reliable and robust measure of attribute importance.

With Ensemble Prediction, attribute importance can be defined as the increase in the average of squared prediction error when replacing the attribute in the model with a random variable. In other words, a variable's importance is its contribution to predictive accuracy. The particular approach that we investigate here is known as Random Forest[1] and proceeds as follows:

1. Randomly sample (with replacement) a training data set from the full data set.

2. Randomly sample a subset of predictor variables from the potential set of predictor variables. From the random subset, choose that predictor variable and its value that splits the data so as to maximize prediction success outside of the sample selected in Step 1.

3. Repeat Steps 1 and 2 multiple times (e.g., 1000 times).

1 **Random Forest is a technique created and written in Fortran by Leo Breiman of UC Berkeley and Adele Cutler of Utah State University. The software was later converted to the R Language by Andy Liaw and Matthew Wiener of Merck Research Laboratories.**

4. For each regression within each training data set, calculate a prediction error using the validation data set (i.e., observations excluded from the training data set).

5. For each training data set, calculate variable importance as the average increase in mean squared error (MSE) of prediction when replacing the attribute in the model with a random variable.

Before we proceed, we will provide a brief comparison of selected techniques for measuring the importance of key drivers of the purchase or prescribing decision. This comparison helps to explain the advantages of Ensemble Prediction.

Advantages and Disadvantages of Selected Approaches in Measuring Strength of Key Drivers

The simplest approach is to use correlation analysis to determine the strength of association between a brand's overall market performance and the perceived performance of the brand on separate attributes. Typically, attribute ratings from a survey provide the data. The technique is easy to execute, but it does not discriminate well between the most important attributes and less meaningful attributes that may only appear important.

A second technique is multivariate regression using the same type of ratings data from survey questions. The regression approach attempts to explain overall market performance as a function of the ratings on separate attributes. This approach is superior to correlation analysis, from a theoretical point of view, since overall market performance is indeed explained simultaneously by many attributes. However, extreme correlation among predictor attributes often causes aberrant results. For example, it might appear that only two or three predictor attributes have a positive impact on overall market performance of the brand. Other predictor attributes may be important influencers of overall market performance, but the correlation of predictor variables prohibits us from being able to statistically measure their unique influence.

A third approach is known as MaxDiff Scaling. This approach overcomes the problem of high correlation among predictor variables by avoiding scales totally. In fact, customers do not rate brands at all. Rather, they select from a short list of attributes the ONE that is most important to their purchase decision and the ONE that is least important. By forcing respondents to choose among attributes, we can measure the relative importance of each attribute based on a probability model—the probability of being chosen as MOST important.

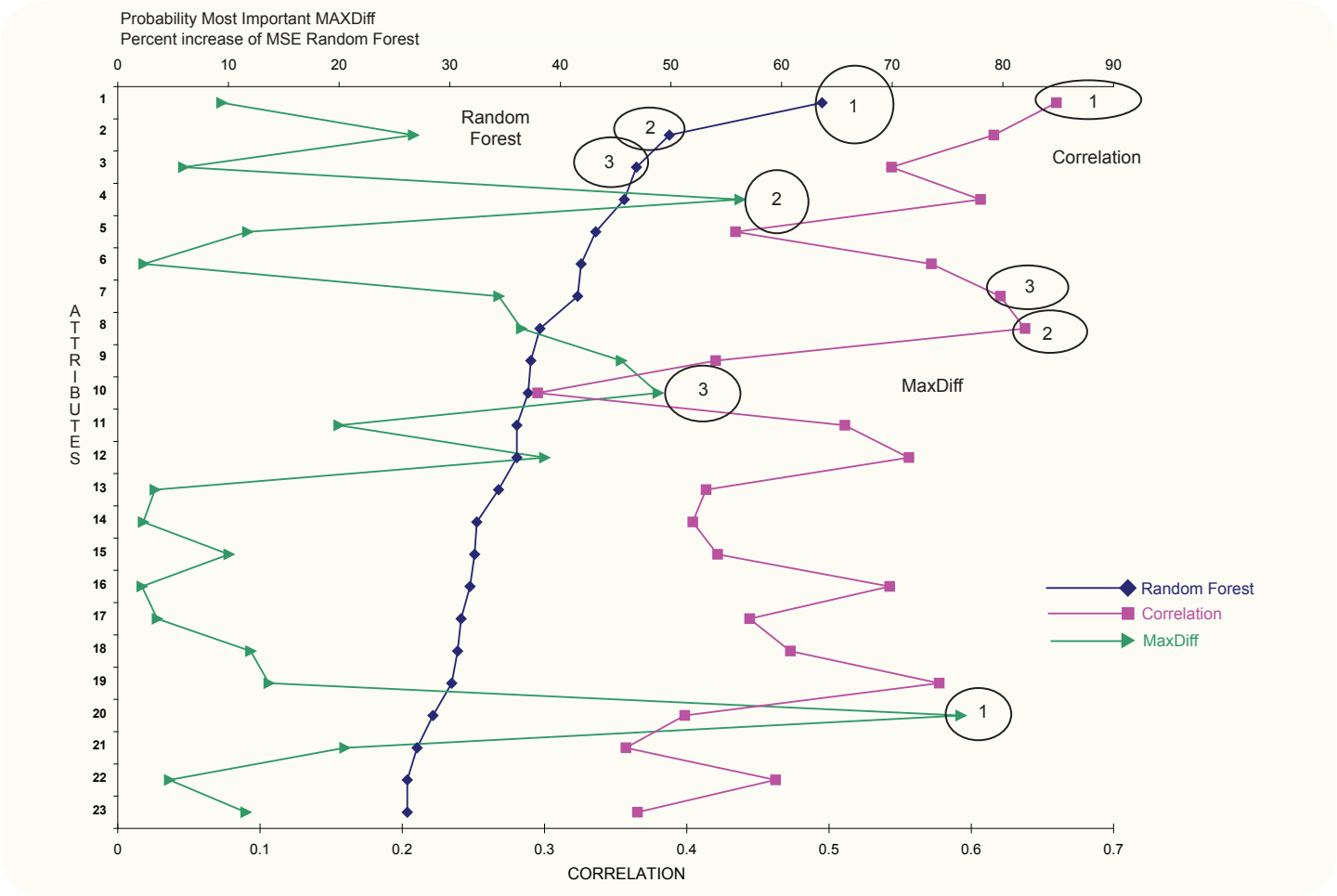
While this technique is newer and avoids the correlation of brand ratings that result from asking customers to rate each brand on multiple attributes, MaxDiff remains a stated importance technique. The drawback of stated importance approaches, in general, is that customers may not be able to articulate why they make their purchase choices. In the case of physicians, relying on them to accurately state the most important factors in their brand choice is particularly problematic. Other factors that physicians are likely less conscious of, or less willing to admit to, may be important to the brand decision.

A fourth approach we investigate, Random Forest, provides an advantage by validating key drivers based on thousands of out-of-sample predictions.

Comparison of Correlation, MaxDiff Scaling, and Random Forest Based on Actual Data

We applied Random Forest regression, correlation analysis, and MaxDiff to a healthcare product category to investigate its potential for use in key driver analysis in marketing research studies. The results provide a visual demonstration of the kind of results we have found in actual applications of Random Forest to key driver analysis.

The following chart represents a pattern of results we have found in applying three of the aforementioned techniques to determine key drivers in a single study. The top-three attributes, ranked in order of importance, are identified for each of the three different techniques. The smoothest curve (black line with diamonds as points) is a plot of importance by attribute, rank-ordered from most to least important, based on Random Forest regression. The



dependent variable being predicted by the regression is an overall rating of brand performance, where survey respondents rate multiple brands with which they are familiar. There are 23 additional attributes used to rate each brand. The measure of the attribute's importance from the Random Forest regression, as explained earlier, is the percent increase of mean squared error (MSE) of prediction (out of sample), when replacing a predictor variable with a random variable.

The purple line (with squares as points) is the plot of correlation for the same variables. It is not a multivariate procedure.

The green line (with triangles as points) is the plot of MaxDiff scores for each attribute. The MaxDiff score is a number from 0 to 100, representing the probability that the attribute would be chosen as the most important if all attributes were to be presented to a respondent.

The following conclusions were made based on our experience with key driver analysis:

1. Correlation and Random Forest usually, as in this example, identify the same top key driver. This is understandable as both correlation and Random Forest regression are "derived importance" measures. By "derived importance" we mean that the measure of importance is the result of "statistically" determining the relationship between a brand's performance on a particular product or service attribute and the brand's overall market outcome.
   In contrast to correlation analysis and Random Forest regression, MaxDiff often points to a completely different top key driver. In the chart, MaxDiff found Attribute 20 to be most important and Attribute 1, the most important attribute based on the derived importance measures, as not very important at all!

2. Based on our internal research in comparing derived importance and MaxDiff attribute importance (as in our example chart), MaxDiff results are fundamentally different from derived importance measures. In particular, those attributes found to be MOST important in MaxDiff are typically obvious—for example, how many physicians would not say that efficacy is most important to prescribing a medication? In addition, MaxDiff will not necessarily find those hidden drivers of the prescribing decision. How many physicians would admit that their relationship with a sales rep is very important in the prescribing decision?

3. Correlation and Random Forest attribute importance generally diverge after the first two or three attributes. In particular, correlation analysis finds a number of attributes to be almost as important as the most important attribute. This finding is consistent with the high degree of correlation of the attribute ratings, causing the analyst to be unable to discriminate between what is more and what is less important.

4. On the other hand, Random Forest regression does appear to distinguish the relative importance of correlated variables.

Implications for Marketing Research

Based on our internal research, we highly recommend the use of Random Forest as a new tool in key driver analysis for marketing research studies. We suggest that bivariate correlations be generally avoided when determining key drivers of purchase or prescribing decisions, since they do not discriminate well between the most important and least important attributes. Given our experience that MaxDiff scaling will often fail to identify attributes that are unacknowledged or "silent" drivers, we suggest caution in its use for key driver analysis.

We recommend Random Forest regression for key driver analysis based on the following reasons:

- A multivariate approach is methodologically superior to a bivariate approach such as correlation analysis.
- Contribution to out-of-sample prediction success is clearly a stronger criterion upon which to base importance, as this type of cross-validation confirms that the results are truly predictive of market outcomes with a different set of customers than those that were interviewed.

About the Author

John Colias s a Senior Vice President of Advanced Analytics at Decision Analyst. The author may be reached by email at *jcolias@decisionanalyst.com* or by phone at *1-800-262-5974* or *1-817-640-6166*.

Decision Analyst is a global marketing research and analytical consulting firm. The company specializes in advertising testing, strategy research, new products research, and advanced modeling for marketing-decision optimization.

Decision Analyst
strategic research ■ analytics ■ modeling ■ optimization

604 Avenue H East ■ Arlington, TX 76011-3100, USA
1.817.640.6166 or 1.800. ANALYSIS ■ www.decisionanalyst.com