A FAIR ISAAC WHITE PAPER

Using Adaptive Random Trees (ART) for optimal scorecard segmentation

By Chris Ralph
Analytic Science Director
April 2006

Summary

Segmented systems of models are widely recognized as a highly effective way of boosting predictive power; however, the traditional process for building these systems is often considered too laborious to justify the effort. Fair Isaac uses Adaptive Random Trees (ART) technology to create segmented models that predict outcomes with great precision. ART uses a powerful global procedure that looks at the entire terrain before determining the best segmentation scheme or “tree,” rather than local “branch by branch” decisions, as in the traditional manual techniques. Rooted in advances in genetic algorithms, ART automates the process of searching thousands of combinations of segmentations and models to find the best segmented model system.
Table of Contents

Segmented Scorecard Development ............................................................................................................ 1
  Traditional approach to segmented scorecard development ................................................................. 2
  ART approach to segmented scorecard development ........................................................................... 2
How ART works ........................................................................................................................................... 3
  ART in practice ......................................................................................................................................... 5
  ART availability ......................................................................................................................................... 5
About Fair Isaac ............................................................................................................................................ 6
Segmented Scorecard Development

The process of predictive model development requires the identification of features in the data that exhibit clear and interpretable patterns with respect to the target variable of interest. Consider, for example, an analyst charged with the task of ranking a population of prospective customers by the likelihood that each will respond to an offer for a new low interest credit card. One of the key predictors of responsiveness to this offer is “age”, which exhibits the following pattern:

[Graph showing Probability of Response vs. Age]

In other words, when the entire population is ranked by age, the analysts notices that younger prospects are more likely to respond, and older prospects are less likely to respond. This discovery could be leveraged to accomplish the task at hand by including age in a predictive model, or scorecard, such that younger prospects get a higher score (i.e., indicating higher response likelihood) and older prospects get a lower score. To build a good scorecard, the analyst needs to continue searching through the data to find other predictors, in addition to age, that lead to other useful patterns with respect to the target that can also be leveraged.

During the course of this investigation, however, the analyst also notices that not all younger prospects are highly likely to respond to this particular offer. In fact, the nice, intuitive response pattern shown above begins to look quite different when the population is split into sub-segments representing different levels of income.

[Graph showing Probability of Response vs. Age (Low income sub-segment)]

[Graph showing Probability of Response vs. Age (High income sub-segment)]
By splitting the population into sub-segments by income, the analyst is able to identify an important fact. It turns out that the relation of response to age is much stronger in the low income sub-segment when compared to the overall population. In fact, it appears that age becomes almost useless as a predictor of response when income is high.

Rather than build a single scorecard for the entire population, this discovery leads the analyst to build a separate scorecard in each of these sub-populations created by splitting on income, with age included as a predictor in the low income scorecard but not in the high income scorecard. As the analyst continues with her analysis in these two sub-populations, she discovers some additional variables that work well in the high-income group that are not predictive in the low-income group, and vice-versa. This leads her to develop a segmented system of models that generates much more accurate estimates of responsiveness than can be accomplished with a single, global model.

**Traditional approach to segmented scorecard development**

The scenario outlined above represents a very common situation faced by analysts who are trying to build the best possible predictions across a wide range of modeling problems. Segmented systems of models are widely recognized as a highly effective way of boosting predictive power, however the traditional process to build these systems is often considered much too laborious to justify the effort.

In the absence of an automated approach, most analysts use a manual or greedy search process. Typically, this involves starting with the entire population and manually selecting segmentation variables. Such as income in the example above; splitting thresholds within those variables one at a time; and building scorecards in each resulting sub-population. The winning split becomes the one that results in the greatest overall boost in predictive power. The process is then repeated in each of the chosen new sub-populations until either predictive power stops improving, or the analysts run out of time or energy.

Fair Isaac has been a worldwide leader in the development of highly effective and robust segmented scorecard systems for more than 20 years. This is due largely to our ability to develop unique technologies that help to streamline the search process and identify the best combination of splitters and split points to define an overall segmentation tree. Our scorecard technology will construct robust predictive models in the leaf nodes of those trees. However, until recently, these powerful segmentation tools still suffered from the pitfalls of a greedy search process. The term greedy stems from the fact that each splitting decision is made using only information from the sub-population that is currently being split. Although we had been able to develop procedures and tools that ensured the selected segments led to robust and effective solutions, this traditional process was inherently near-sighted, because it was incapable of calculating the effects of local decisions—i.e., those regarding the current sub-population—on downstream segments. In other words, a splitting decision made near the top of a decision tree could lead to subsequent sub-optimal splits that adversely affect the overall system of scorecards.

**ART approach to segmented scorecard development**

Adaptive Random Trees is a highly innovative technology that not only addresses the greedy search problem, but also provides an intelligent, automated process to find powerful segmentation systems in a small fraction of the time required for the manual approach described above. ART is designed to work with binary target variables. With a little bit of clever data pre-processing, ART can also be used on continuous outcome targets. Rather than search through sets of splitters and split points in successive order to build a segmentation tree one level at a time, ART works by evolving an entire population of fully defined trees over time towards increasing overall predictive power. Controlling this search is a proprietary genetic algorithm, which works by selecting sets of the most effective trees and swapping information—i.e., splitters and split points—between them to create successive generations of new trees.
Over time, as this process continues, the best overall combination of splitters and split points emerges from this population, and the analyst is provided with the segmentation logic for a system that has been optimized with a global objective function that measures overall predictive power.

**How ART works**

The first step in developing a segmented scorecard system using ART is to develop an initial “parent” scorecard on the entire population. Not only does this provide a benchmark against which the segmented systems can be compared, it is also used to define a set of binned predictor (independent) variables that will appear in the segment-level models in every leaf node of each tree that ART considers. Instead of calculating a single coefficient for each predictor, as in a linear regression model, the scorecards work by defining a set of carefully chosen bins within each predictor variable and associating a score weight to each of those bins. This approach enables the resulting scorecard to take full advantage of the underlying patterns inside each predictor with respect to the target, including any non-linear patterns that are often missed by traditional regression techniques.

Although every leaf-node scorecard uses a common set of binned predictors, the resulting models are specifically predictive for that sub-population because they contain segment-specific bin-level score weights. In ART, those weights are set to the weight of evidence values calculated within the bins.

\[
WoE_i = \log_e \left( \frac{P(bin_i | outcome_i)}{P(bin_i | outcome_2)} \right)
\]

So, every observation falls into one of the bins of each predictor variable in the leaf node scorecards, and the overall score for each observation is simply the sum of the WOE values from those bins, normalized to a common scale. Scorecards that use WOE values based on these simple bin counts are referred to as naïve Bayes models. These models are used in ART because not only are the WOE values highly interpretive and robust when used as score weights, they can also be calculated very quickly using simple counts. This is particularly useful in ART, where many thousands of score weights are calculated in a typical run.

Once the parent scorecard is created and the binned predictor variables are defined, the next step is to use ART to identify a set of binned candidate splitter variables. These can consist of all the variables considered during the construction of the parent scorecard, including those that made it into the model for the overall population. Given the binned predictors and the binned candidate splitters, an ART run can be launched.

An ART run consists of a number of generations, each of which contains a fixed number of trees. In the initial population (i.e., generation #1), the trees are randomly created from the user-defined list of splitters and split points, along with a set of parameters that constrain the total number of leaves. In each leaf of each tree, ART constructs a naïve Bayes scorecard using the binned predictors that appeared in the parent scorecard. A score is generated for every observation by summing the WOE values associated with the bins in the scorecard of the leaf that observation fell into.
Given the scores for each observation, the measure of predictive power used by ART to evaluate the effectiveness of each system of scorecards is Divergence, which is calculated as follows:

\[
\text{Divergence} = \int \left( P(S \mid \text{outcome}_1) - P(S \mid \text{outcome}_2) \right) \log_e \left( \frac{P(S \mid \text{outcome}_1)}{P(S \mid \text{outcome}_2)} \right) ds
\]

Divergence simply measures the ability of the scores from a given tree to separate the two outcome classes, where higher values indicate a greater degree of predictive power. In ART, this Divergence value calculated from the system of scorecards appearing in the leaf nodes of the segmentation trees serves as the fitness value that drives the genetic algorithm search.

After Divergence has been calculated the genetic algorithm takes over. Below is a snapshot of the genetic algorithm search process in ART.

The genetic algorithm first ranks the trees by Divergence and identifies an elite set, which are the best ‘N’ trees by this measure of fitness. This elite set is automatically promoted to the next generation of trees. This is important because it serves as the memory of the search algorithm, always retaining the best solutions found so far.

Next, the genetic algorithm uses a process called fitness proportionate selection to choose pairs of parent trees that will be mated to create new child trees. Fitness proportionate selection leverages the Darwinian concept of survival of the fittest, whereby trees with higher Divergence are more likely to be selected for mating. The process itself is akin to a roulette wheel, where each tree is assigned a slot on the wheel, and the relative sizes of the slots is proportional to the relative divergence of the trees.

Following fitness proportionate selection, the selected pairs of parent trees are then mated to create new child trees that will comprise the subsequent generation. Mating is accomplished by randomly selecting a node in the mother tree and replacing its entire sub-tree with the sub-tree selected from the father tree. This branch swapping procedure is then checked for logical consistency, and any empty or illegitimate nodes are pruned back up to the lowest valid parent node.
The final step in the creation of the next generation of trees is called mutation. It involves randomly choosing a subset of the newly created child trees and replacing one or more of its splitters and split points with one selected from the user-defined list of possible candidate splitters. This process of mutation helps to maintain diversity in the search, and in many applications, this diversity is the key to ensuring a near-optimal final solution.

This entire process is repeated over successive generations of trees until the values of Divergence stop improving on a holdout data sample. At that point, ART will return to the analyst the segmentation logic and leaf-node scorecard details of the very best overall set of trees.

**ART in practice**

Adaptive Random Trees is a breakthrough technology that has enabled the application of the segmented scorecard methodology across a wide range of applications. As the inventor of this technology, Fair Isaac has benefited from the massive scalability enhancements enabled by this unique and highly focused automated data mining tool. Our clients have also benefited from much more effective segmentation solutions enabled by a smarter tree search technology, delivered with greatly reduced turn-around times. As we continue to expand our genetic algorithm-based suite of predictive modeling power tools, we look forward to continuing our work with our clients and partners to deliver more world-class decision management solutions.

**ART availability**

Segmentation ART is now available as an optional module for the Model Builder scorecard platform. Designed by seasoned modelers, Segmentation ART overcomes the challenges that have made segmented model systems impractical for many analytic development projects. Segmentation ART technology leverages the capabilities of Model Builder scorecard for importing data, visualizing and exploring predictive patterns, defining predictive variables, creating models, evaluating their quality and swiftly deploying predictive analytics.

By automating the process of identifying optimal segmentation, Segmentation ART delivers substantial benefits to your model development process. The ART process results in more elegant trees. These trees often reduce the number of models required, and dramatically cut development, deployment and maintenance costs. ART segmentation helps you boost the precision of your models, enhance the speed of the model development process, improve efficiency and reduce costs.
About Fair Isaac

Fair Isaac (NYSE:FIC) makes decisions smarter. As the world leader in decision management solutions driven by advanced analytics, Fair Isaac unlocks value for people, businesses and industries. Companies worldwide use Fair Isaac technology to make billions of faster, more profitable decisions a year in credit management, marketing, fraud, collections, bill payment and other areas. The world’s leading banks and credit card issuers rely on Fair Isaac technology, as do insurers, retailers, telecommunications providers, healthcare organizations and government agencies. Founded in 1956, Fair Isaac serves thousands of clients through offices in nine countries, and helps millions of individuals improve their credit health through the www.myfico.com website. Visit Fair Isaac online at www.fairisaac.com.