# Ensemble Modeling

## Get More Out of Your Predictive Modeling

Gary Robinson
COO | Lityx LLC

Just as an ensemble is a collection of parts that contribute to a single effect, ensemble modeling is multiple models used together to outperform any single one of the contributing models. It is based on the philosophy that, "Together, we are better." Do you recall the Netflix Prize from a few years ago, when Netflix offered $1 million to anyone who could improve their movie recommendation algorithm by 10 percent? After three years of competition, the prize was awarded to a team that formed from other teams and combined their thinking and models into a single model. Their approach really shined a light on the power of ensemble modeling. Ensemble modeling can get very complex, but you can benefit from this approach by beginning very simply.

We all know that models use current and past information about customers and prospects to predict a future behavior such as buying, opening, responding, churning, charging off, etc. There is a wide variety of appropriate techniques to choose from when you build a predictive model but, generally, they produce similarly effective results. The majority of modelers use a well-known, preferred technique for most business problems, but other techniques might get used, depending on the modeler's experience and personal preference.

For example, logistic regression is the technique most often used to solve business problems that can be viewed in a binary manner – things such as click (yes/no), churn, (yes/no), buy product A vs. B, etc. However, other techniques can solve these types of problems just as effectively, most notably decision trees such as CART and CHAID, neural networks, and machine learning algorithms like SVM, to name a few.

What is interesting is that while the different techniques produce similarly effective results, they go about their jobs in very different ways. For example a regression based approach seeks to find the set of predictors that collectively minimizes the sum of the squared differences between actuals and predicted. A decision tree on the other hand chooses predictors one at a time independently selecting the next one that provides the biggest statistically significant split in performance. In some situations one technique is more appropriate than another, but often any one of a handful of techniques can be appropriate. On their own, each technique can provide the business a lift that drives thousands or millions of dollars in performance benefits. But why do we have to choose only one? Like in the Netflix challenge, when you have many points of view you not only have a better chance of finding a solution, but also are often able to come up with an even better solution using parts of multiple ideas. This, in essence, is what an ensemble model is. Combining multiple – or an ensemble of – techniques to arrive at a solution that outperforms any of the solutions derived from a single technique.

### Building an Ensemble Model

The approach I use for building ensemble models is first to build a variety of models, each using a different technique but all trying to predict the same business outcome. Then use the output scores from each of the models as predictors for a second round of modeling, this time using only your preferred modeling technique. Consider the following example.

For a typical classification model, say response, we might prefer the use of logistic regression on behavioral and demographic data to best predict responders from non-responders. The validation results in **Figure 1** show a nice rank order of the deciles, with decile 1 most responsive, at 27 percent, and decile 10 least responsive, at 4 percent.

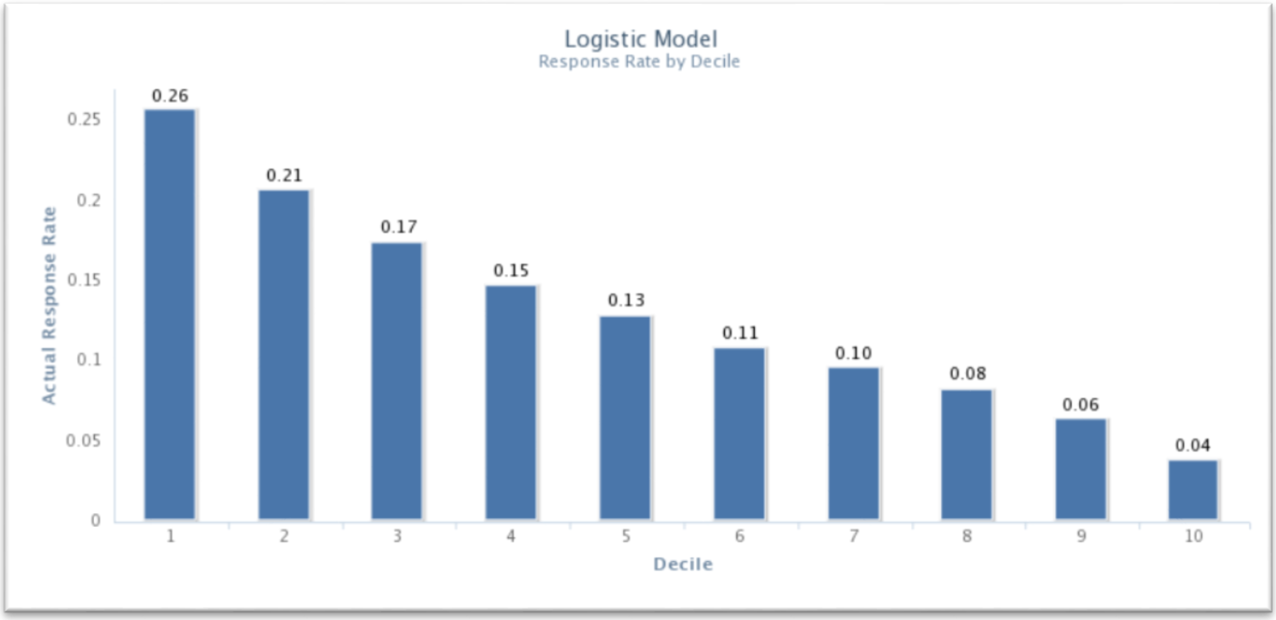

**Figure 1**. *Actual response rates from a hold-out sample (not used in the model development) split out by model decile where decile 1 is expected to have the best performance descending to decile 10 expected to have the worst performance.*

Now let's test the use of a decision-tree approach and build a CHAID model. In **Figure 2**, we see the logistic and CHAID model validation performance side by side, and that the CHAID model does not rank order as well as the logistic model, going from 23 percent to 5 percent.

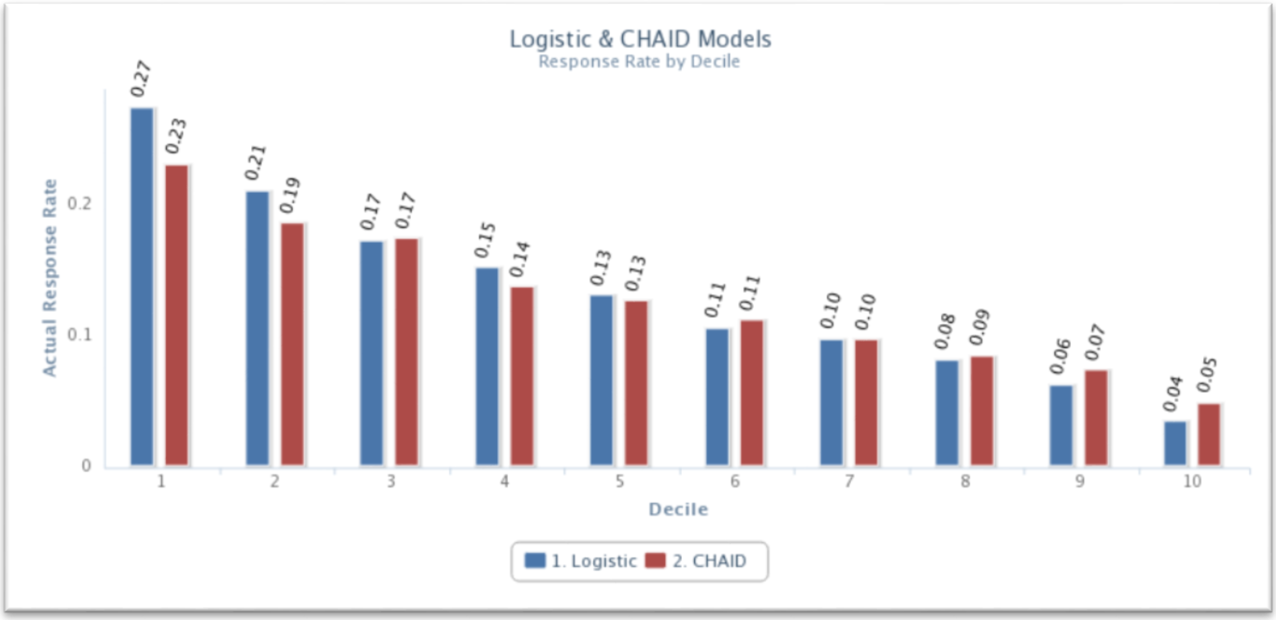

**Figure 2.** *Use of a CHAID decision tree on this example data to separate response rates into ten deciles shows not to be as effective an approach as logistic regression. The CHAID decile response rate range (best performance subtract worst) is not as good as the Logistic response rate range.*

Next, we will try an ensemble model built using a logistic regression approach on only the predicted scores of the logistic and CHAID models as inputs. **Figure 3** validation results show us that by combining the logistic and CHAID approaches, we are able to get a slightly improved model over the initial logistic model only, with the deciles of the ensemble model going from 28 percent to 3 percent.

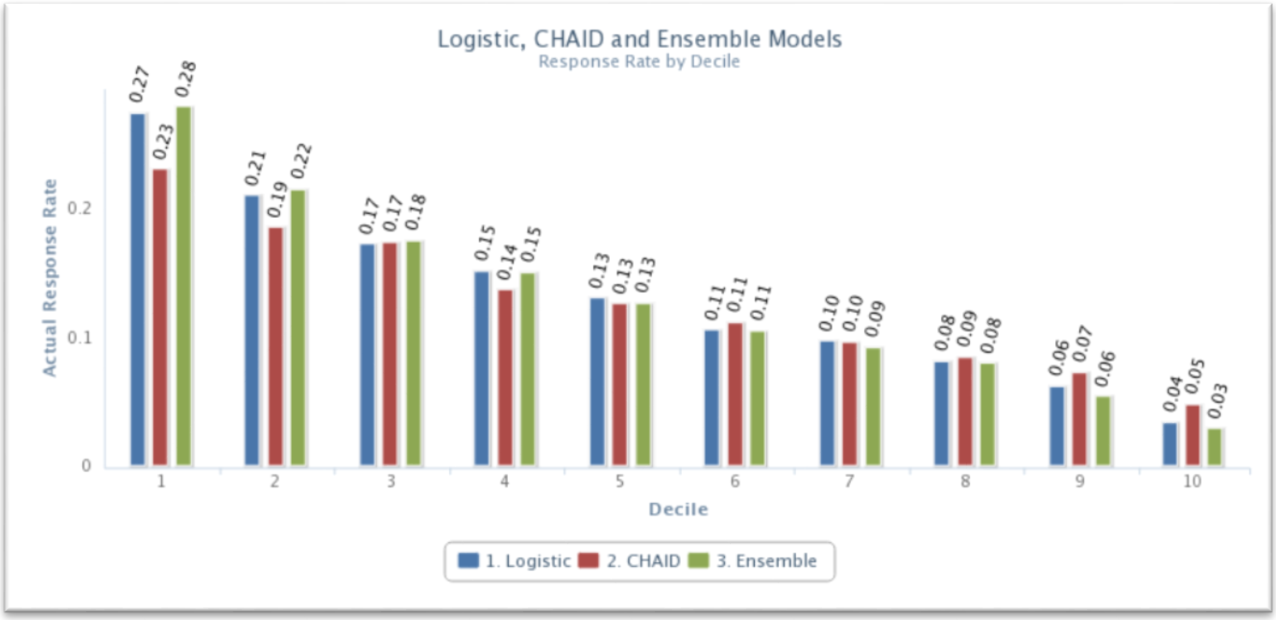

**Figure 3.** *The Ensemble model produces a greater range in decile response rates than the Logistic (.25 vs .23) and the CHAID (.25 vs .18).*

An ensemble model is not guaranteed to outperform its component models, but when it does it is because there is some incremental insight that one approach was able to capture that the other was not. Sort of like filling in the small cracks of missing explanation that a single model approach leaves behind. Small improvements in large volumes can generate substantial gains. Ensemble modeling approaches can become very involved, but hopefully you see the approach presented here is simple enough to warrant a review the next time you are looking to try and get more out of your predictive modeling!

---

*Gary Robinson is Chief Operating Officer of Lityx. He is a 20-year veteran of marketing analytics and CRM, having most recently worked at Merkle Inc., where he was Vice President in the Quantitative Marketing Group overseeing analytic engagements for some of Merkle's top clients. Prior to Merkle, Gary was an SVP at Bank of America, where he held various positions including marketing director and head of database marketing, as well as running strategy, analytics, and list procurement.*

*Gary has a master's degree in Statistics and a bachelor's degree in Psychology from Arizona State University.*