



"turning data into dollars"

Tom's Ten Data Tips – December 2009

Predictive Modeling

Predictive modeling is where data miners' most visible "claim to fame" lies. Advances in software usability, computing power, and algorithms have brought this capability within the realm of modern business. Whether it is direct marketing, credit scoring, fraud detection, or forecasting, predictive modeling (by employing data mining) has proven its worth.

Predictive modeling is foremost empirical work. You need a modicum of database knowledge, and interest in statistics and methodology, and with many tools some programming knowledge. Because defining the modeling requirements is so important, we have found that excellent communications skills are of utmost importance to be effective. Don't just build great models, make sure they are "the right" ones... (see e.g.: tip# 5)

1. Predicting Is Classifying The Past

What we commonly refer to as predictive modeling, isn't really "predicting" at all. What we refer to as a "predictive model" is based on identifying patterns in the *past*, and then finding lookalikes of those instances in the *present*. We make predictions under the strict assumption (and this is a big assumption) that the future will look like the past. So what we actually do is classify the present with reference to the past.

All so-called predictive models that *restrict* "predictions" to conditions that have previously occurred fall into this category. Making predictions *outside* ranges that were incorporated during model development tends to invalidate results (for these techniques). This, incidentally, is where disaster loomed in the credit crisis: conditions occurred that had never been observed before, yet people continued to rely on these "predictive" models.

Models that *can* possibly deal with states that have never occurred, yet, are for instance system dynamics models. For ease of reading we'll continue to use the term "predictive modeling" the way it is commonly used, as if it really *were* a prediction.

2. Monetize *Every Model, In Any Way Possible*

You should bend over backwards to estimate the (monetary) value of a model in terms of one-dimensional benefits to the business, preferably dollars (or Euro's, GBP, or whatever currency you have). This helps you to "sell" your models better to the business (the financial argument really *does* work). But it also helps you to assess better how to allocate your time (see also tip# 5).

One of the reasons why monetizing results can be difficult is because many organizations that (begin to) work towards accountable marketing practices experience hick-ups and resistance while making this transition. A sure-fire symptom is an inability to calculate the value of a product or the (marginal) cost of contacts. Maybe someone, somewhere does not want to be reminded that the actual cost of an outbound telemarketing call is \$15-\$50 (a fair guess for *short* sales calls, based on our experience), after you factor in *all* the costs. These numbers might be upsetting to some. Similar with calculating customer profitability, which happens to be genuinely difficult. See a previous newsletter on customer profitability.

3. Some Predictions Are Made To Be Refuted

In many cases the ulterior motive behind prediction isn't just *foresight* but rather *control* of the future. A revenue forecast that doesn't meet business objectives isn't merely "accepted." Instead, targets are set to do better than is to be expected (see a previous newsletter on forecasting, notably tips# 1 & 6). For a collections department, a default prediction *will* help in planning workloads, but that is not the main reason to calculate these scores. More importantly, these accounts will (or should...) be actively managed to *avoid* defaulting.

What these examples have in common is that predictions *are* accepted as valid, yet they drive interventions that are meant to change an imminent (and undesirable) reality. So we do not wait for "nature" to take its course. Instead, we intervene, attempt to change that reality. Needless to say that when such interventions *are* successful, this will trigger a need for new and adjusted predictive models.

4. Make Sure Your Predictions Arrive "On Time"

Whenever a prediction is made that is supposed to be acted on (as in tip# 3), you need to factor in the required deployment time. This impact is not always understood. "Technically", it implies that you need to predict further out into the future. Otherwise predictions would

not be available when you need them, and/or still in time to act upon them. There is little use for predictions about which accounts are bound to default if these scores become available when it is too late to take corrective action.

Take for example churn of cell phone contracts. It is well known that the closer you get to the expiration date of your contract, the higher your chances of canceling the subscription. From the calling behavior itself it is not always obvious who will cancel, and who will not, nor when. You need to be able to make that prediction (with sufficient accuracy) *before* the actual decision to cancel has been made. After that point in time you might be able to predict quite well, but will be ineffective with your retention interventions (you wind up in a “win-back” scenario). At the same time, the prediction needs to be accurate enough to offset any benefits you offer to churn candidates who would have stayed anyway, since these constitute a net loss that needs to be offset by success in “saving” likely cancelations.

5. Paint With A Broad Brush If Misclassification Costs Are Low

When is a predictive model good enough? When should you go the extra mile, and when should you say: “good enough for government work”? At a micro level, it may seem well worth your time to invest a few additional days to improve lift (=predictive accuracy) of the model. Additional revenues from the model easily offset your hourly wage. Still, you need to take a look at the bigger picture.

There are almost always more opportunities for applying models. The talent needed to assess where a model could add value, and “making it work” is often the scarcest resource. So don’t squander it fine tuning a model that already works very well for the business. The question is not “is it worth my time?” but rather “where is my time best spent?”

6. Explaining Predictive Models Requires (At Least) Working Knowledge Of Statistics

Because predictions are never perfect (or else they stink: are guaranteed to be *wrong*), and you will often need to explain what can and cannot be expected, you need to be *very* comfortable explaining statistics in lay men’s terms. Using simple and intuitive analogies, and filling them in with “real life” numbers helps people grasp abstract concepts like lift curves and conversion probabilities. In part this has to do with managing expectations, and in part with educating business “consumers” who could take advantage of using predictive models.

A perfect prediction would nor could ever be possible. It would imply a deterministic world: a given set of input conditions would then *always* lead to a particular outcome. Most people would like to believe we are (more or less) free individuals, and we can “choose” our course of action. Perfect, or near perfect predictions would refute this view of mankind.

7. The Quality Of A Model Can *Never* Be Captured In One Single Number

Every data mining algorithm has an idiosyncratic measure to assess how accurate a prediction is. R^2 for regression, H-L for logistic regression, χ^2 , Gini, entropy (& others) for decision trees, etc. These numbers are useless for comparison *across* algorithms. For that purpose, lift is the most commonly used measure. Lift indicates how much better the predictive model performs over a random guess.

What is not well understood is that the predictive quality of a model can fundamentally *never* be captured in one single number. For one thing, this has to do with the measure of uncertainty “surrounding” the prediction. Also, when we talk about a prediction, this implies generalizing to *other* data sets than the mining set (training-, test-, and evaluation set being random samples from the *same* mining set, of course). How well a model generalizes in the face of population drift has to do with multi-variate representativeness, one of the frontiers of modern statistics.

8. There Are Three Ways To Improve Your Model

Many data miners like to tinker with their models, tuning them to improve predictive accuracy. If misclassification costs are high (see also tip# 2), it is generally worth the effort to squeeze the very last drop out of your data. This can be done in one of three ways: selecting other (better) variables to include in your mining set, preparing your data better, or selecting/tuning the algorithm you use.

Selecting better variables is almost like “cheating”, because you look outside the boundaries of the mining set. For a dramatic improvement, that’s where you generally need to look. Preparing the data better means imputing substitutes for missings, optimal binning of continuous variables and mapping of categoricals. It’s good practice to habitually compare a range of algorithms. After you have settled on the most promising one, tune it to arrive at your final model.

9. Data Preparation Makes A (Big) Difference

Because of practical/logistical constraints with regards to assembling datasets and deployment, the available data set is (usually) more or less fixed. So finding new or better variables tends to be more of a long-cyclical effort. Highly worthwhile, but typically meager short-term results. Selecting a “better” algorithm should be common practice, and can sometimes (though not often) give surprising results. Tuning the algorithm may well be attractive, and within the comfort zone of the data miner, but probably ranks lowest in terms of effect.

For the highest impact (certainly short-term) carefully preparing the data usually reaps the most rewards. The (raw) data are already available, so no one else needs to be involved (it has low impact on the organization). How to treat missings, how to map and/or transform categoricals, etc., can have a huge impact. These are (relatively) easy quick wins. Needless to say, the exact same operations need to be performed on all (two or three) sections of the mining set, *and* the data you will deploy to.

10. Monitor, Monitor, Monitor

The most common “error” in predictive modeling is omitting to *constantly* look at the past. The longitudinal trace of how you’re predictions stand the test of time hosts tremendous value. It is also a necessary safeguard against bad predictions. Unfortunately, tracking how predictions evolve (degrade) as reality begins to drift is often an afterthought. Some companies neglect this altogether, and some don’t know how to do it (properly).

Some variables retain their predictive power, and some decay. Domain knowledge on your data is the coinage of your realm as predictive modeler. It will also help you choose between a great model that might not have quite the half-life of another one with slightly lower lift.