



"turning data into dollars"

Tom's Ten Data Tips – August 2006

Data Mining Models

What is a model? A model is a purposeful simplification of reality. Models can take on many forms. A built-to-scale look alike, a mathematical equation, a spreadsheet, or a person, a scene, and many other forms. In all cases, the model uses only *part* of reality, that's why it's a simplification. And in all cases, the way one reduces the complexity of real life, is chosen with a *purpose*. The purpose is to focus on particular characteristics, at the expense of losing extraneous detail.

If you ask my son, Carmen Elektra is the ultimate model. She replaces an image of women in general, and embodies a particular attractive one at that. A model for a wind tunnel, may look like the real car, at least the outside, but doesn't need an engine, brakes, real tires, etc. The purpose is to focus on aerodynamics, so this model *only* needs to have an identical outside shape.

Data Mining models, reduce intricate relations in data. They're a simplified representation of characteristic patterns in data. This can be for 2 reasons. Either to predict or describe mechanics, e.g. "what application form characteristics are indicative of a future default credit card applicant?". Or secondly, to give insight in complex, high dimensional patterns. An example of the latter could be a customer segmentation. Based on clustering similar patterns of database attributes one defines groups like: high income/ high spending/ need for credit, low income/ need for credit, high income/ frugal/ no need for credit, etc.

1. A Predictive Model Relies On The Future Being Like The Past

As Yogi Berra said: "Predicting is hard, especially when it's about the future". The same holds for data mining. What is commonly referred to as "predictive modeling", is in essence a *classification* task.

Based on the (big) assumption that the future will resemble the past, we classify future occurrences for their similarity with past cases. Then we 'predict' they will behave like past look-alikes.

2. Even A 'Purely' Predictive Model Should Always (Be) Explain(ed)

Predictive models are generally used to provide scores (likelihood to churn) or decisions (accept yes/no). Regardless, they should always be accompanied by explanations that give insight in the model. This is for two reasons:

1. buy-in from business stakeholders to act on predictions is of eminent importance, and gains from understanding
2. peculiarities in data *do* sometimes arise, and may become obvious from the model's explanation

3. It's Not About The Model, But The Results It Generates

Models are developed for a purpose. All too often, data miners fall in love with their own methodology (or algorithms). Nobody cares. Clients (not customers) who should benefit from using a model are interested in only one thing: "What's in it for me?"

Therefore, the single most important thing on a data miner's mind should be: "How do I communicate the benefits of using this model to my client?" This calls for patience, persistence, and *the ability to explain in business terms* how using the model will affect the company's bottom line. Practice explaining this to your grandmother, and you will come a long way towards becoming effective.

4. How Do You Measure The 'Success' Of A Model?

There are really two answers to this question. An important and simple one, and an academic and wildly complex one. What counts the most is the result in business terms. This can range from percentage of response to a direct marketing campaign, number of fraudulent claims intercepted, average sale per lead, likelihood of churn, etc.

The academic issue is how to determine the improvement a model gives over the best alternative course of business action. This turns out to be an intriguing, ill understood question. This is a frontier of future scientific study, and mathematical theory. Bias-Variance Decomposition is one of those mathematical frontiers.

5. A Model Predicts Only As Good As The Data That Go In To It

The old "Garbage In, Garbage Out" (GiGo), is hackneyed but true (unfortunately). But there is more to this topic. Across a broad range of industries, channels, products, and settings we have found a

common pattern. Input (predictive) variables can be ordered from transactional to demographic. From transient and volatile to stable.

In general, transactional variables that relate to (recent) activity hold the most predictive power. Less dynamic variables, like demographics, tend to be weaker predictors. The downside is that model performance (predictive “power”) on the basis of transactional and behavioral variables usually degrades faster over time. Therefore such models need to be updated or rebuilt more often.

6. Models Need To Be Monitored For Performance Degradence

It is adamant to always, always follow up model deployment by reviewing its effectiveness. Failing to do so, should be likened to driving a car with blinders on. Reckless.

To monitor how a model keeps performing over time, you check whether the prediction as generated by the model, matches the patterns of response when deployed in real life. Although no rocket science, this can be tricky to accomplish in practice.

7. Classification Accuracy Is *Not* A Sufficient Indicator Of Model Quality

Contrary to common belief, even among data miners, no single number of classification accuracy (R^2 , Gini-coefficient, lift, etc.) is valid to quantify model quality. The reason behind this has nothing to do with the model *itself*, but rather with the fact that a model derives its quality from being applied.

The quality of model predictions calls for at least two numbers: one number to indicate accuracy of prediction (these are commonly the *only* numbers supplied), and another number to reflect its generalizability. The latter indicates resilience to changing multi-variate distributions, the degree to which the model will hold up as reality changes very slowly. Hence, it’s measured by the multi-variate representativeness of the input variables in the final model.

8. Exploratory Models Are As Good As the Insight They Give

There are many reasons why you want to give insight in the relations found in the data. In all cases, the purpose is to make a large amount of data and exponential number of relations palatable. You knowingly

ignore detail and point to “interesting” and potentially actionable highlights.

The key here is, as Einstein pointed out already, to have a model that is as simple as possible, but not too simple. It should be as simple as possible in order to impose structure on complexity. At the same time, it shouldn't be too simple so that the image of reality becomes overly distorted.

9. Get A Decent Model Fast, Rather Than A Great One Later

In almost all business settings, it is far more important to get a reasonable model deployed quickly, instead of working to improve it.

This is for three reasons:

1. A working model is making model; a model under construction is not
2. When a model is in place, you have a chance to “learn from experience”, the same holds for even a mild improvement – is it working as expected?
3. The best way to manage models is by getting agile in updating. No better practice than doing it... J

10. Data Mining Models – What's In It For Me?

Who needs data mining models? As the world around us becomes ever more digitized, the number of possible applications abound. And as data mining software has come of age, you don't need a PhD in statistics anymore to operate such applications.

In almost every instance where data can be used to make intelligent decisions, there's a fair chance that models could help. When 40 years ago underwriters were replaced by scorecards (a particular kind of data mining model), nobody could believe that such a simple set of decision rules could be effective. Fortunes have been made by early adopters since then.