NEW JERSEY LAWYER

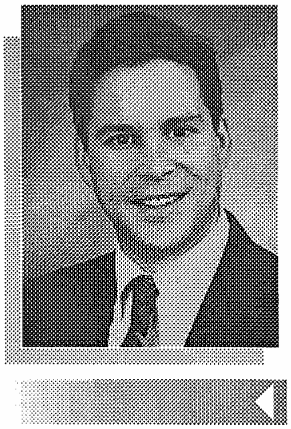
Volume 12, Number 25 The Weekly Newspaper June 30, 2003 $5

'Code the mode': Basics of statistical model development

By Kurt E. Anderson

Statistical models are becoming more and more frequent in all aspects of commerce. Most people are exposed to them in political elections. The major media outlets use statistical models to predict elections winners based on exit polls. Similarly, lenders use statistical models to evaluate applicants' credit risk.

Statistical models now also are being regularly used for many other applications, including everything from evaluating insurability and predicting the success of marketing campaigns to guiding securities transactions for hedge funds. There is no limit to their future application.



Companies frequently rely on independent contractors to develop statistical models rather than maintain in-house staff for this purpose. Unfortunately, contract negotiations for the development of statistical models are frequently stymied by a lack of understanding of the intellectual property issues involved. This article will discuss some of the more befuddling concepts.

The basics

Statistical models, in essence, are complex mathematical algorithms or formulas. They are derived or "built" by analyzing data. Lots of data. Typically, the "predictiveness" (and thus the true value) of a model is a function of the skill of the developer, the quantity of data available to build the model, the quality of the data and the degree to which the model will be applied to a narrowly focused situation.

From a value perspective, statistical relationships and the weighting coefficients associated with them form the two principal components of most statistical models.

Each model also will contain several variables that have statistically significant relationships with the subject the model is formulated to predict. These are the relationships between two or more data elements that are found by the developer to be predictive. Let's use an example very near and dear to my heart: wine. Suppose we were building a statistical model to predict the size of a grape crop for a wine producer. An analysis of the data may reveal that average temperature is statistically associated with the size of the grape crop. If so, we would use this statistical relationship, among others, in our model.

Each statistical relationship will be given a particular weight depending on the predictiveness of that relationship. The predictiveness of the model is primarily a function of the predictiveness of each statistical relationship and the weight given to each within the model. Statistical relationships within a model are given a weight by using a coefficient that gives the statistical relationship greater or lesser influence on the model as a whole. In the example, if the statistical relationships included average temperature, air humidity, type of fertilizer and precipitation, it may be that we would give greater weight to temperature and precipitation and less to air humidity and type of fertilizer.

The vast majority of statistical relationships found in any set of data used to build a model within a given industry generally are known and it is very unusual for a new one to be discovered. Statistical relationships are the tools by which developers build models. However, the composition of statistical relationships within a given model — and especially the weighting coefficients associated with those relationships — can be unique to each model.

Whether all the weighting coefficients are unique may depend on the industry involved. For example, in the credit scoring industry, it is not unusual for all the weighting coefficients in a custom model to be specifically derived. However, in the insurance industry, model developers may use weighting coefficients that are commonly known and consistently used in the industry in combination with weighting coefficients that are specially derived for a particular statistical model.

Statistical models are typically delivered in the form of a written report or software, which then can be used to write the software to automate the calculations performed by the model. Model developers may or may not provide the computer programming services needed to "code the mode."

Kurt E. Anderson is an associate at Giordano, Halleran & Ciesla in Middletown. He represents computer technology-based companies.

Types of models

■ ***Patent.*** In theory, as an algorithm, a statistical model could be subject to patent protection. However, to obtain it, one would be required to disclose the algorithm in the patent application. Since most statistical models are developed to obtain a competitive edge, disclosing the specifics of the model in a patent application would not be a very effective means of protecting the asset.

■ ***Copyright.*** Attempting to apply concepts of copyright gets even more confusing. Copyright protects the expression of an idea, not the idea itself. Since statistical models are primarily utilitarian, there is usually not a great deal of creative expression capable of copyright protection. The contents of the written report or the software containing the model, however, may very well be susceptible to copyright protection. Accordingly, it is important to differentiate between the model itself and the written deliverables or the software used to implement the model.

■ ***Trade secret.*** In most cases, the true value of a statistical model is that no one else knows it. In other words, it is generally protected as a trade secret. Since the law recognizes trade secrets as a property right (similar to a right to ownership of tangible property), trade secret is an effective means of protecting statistical models while being able to exploit their value.

Who owns the model?

But the better question is: Who should own which pieces of the model?

To effectively and fairly determine ownership issues in a statistical model development agreement, it is important to remember the component parts of the deliverables and the role of the developer.

The deliverables in any such agreement will consist of the model itself (composed of statistical relationships or variables and the weighting coefficients); the written report describing the data analysis and the model itself, among other things; and potentially, the software.

■ ***The model itself.*** Since the developer is in the business of building statistical models, the developer should be unwilling to assign all its interest in the model. To do so may preclude the developer from reusing the statistical relationships found within the data. Conversely, the company for whom the model is developed will want to ensure the developer does not give the model to competitors. A reasonable compromise in this situation is for the model to be licensed to the company for which it is developed, provided the composition of the model, taken as a whole and in particular those specially derived weighting coefficients, together constitute trade secrets of the company for whom the model is developed.

From the developer's perspective, the only potential exceptions to this rule would be to ensure that the developer is free to reuse statistical relationships contained in the model and that the developer is free to independently develop even an identical model from data provided by another company for that purpose.

■ ***Written report.*** Model developers may frequently reuse substantial portions of the written material provided to recipients of their models. Accordingly, developers may assert copyright protection in much of the content in the report.

■ ***Software.*** The issues regarding software that implements a statistical model are no different from those associated with any other custom software development agreement. The developer will want to be able to reuse much of the code for subsequent customers. Accordingly, such software should be licensed by the developer to the company for which the model is developed.

Too often, lack of understanding gives way to the approach of parties wanting to own it all because they don't know what is important for them to own. This frustrates companies trying to procure statistical models and reduces the sales cycle times of developers. Understanding the fundamentals of statistical model development is essential to avoiding unnecessary conflicts in negotiating their agreements.

Reprinted with the permission of *New Jersey Lawyer*© June 30, 2003