

Three Trends in Predictive Modeling

Robert Grossman
Open Data Group

September, 2002

1 From Score Cards to Trees

At the most basic, a predictive model takes attributes about a customer or transaction and produces a score. For example, the score may indicate the likelihood of attrition, the likelihood of delinquency, the expected response to an offer, the expected profit, or the lifetime value of a customer.

Predictive modeling for financial services began in earnest with score cards. A score card awards or subtracts points when attributes take certain values (for example, subtract three points if the customer has moved in the past year). Score cards are simple to explain — they are essentially no different than the quizzes present in many teen magazines — but do not have the predictive power of more powerful techniques.

Next in complexity are logistic regression models. The score produced by these models is related in to a linear function of the attributes about the customer or transaction. These types of linear predictors are simple to compute, but often do not capture the complexity of data. For example, linear functions of age, must grow (or shrink) with age, while, for example, the likelihood of delinquency does not vary so simply with age.

More recently, a variety of other models have been used for scoring, such as neural nets and tree-based classifiers. These are examples of nonlinear models. Over the next several years, these should emerge as standard tools. They allow for a more complex relationship between the attributes and the scores and therefore can be used to create more accurate models.

Neural networks and tree-based classifiers are two of the core algorithms that form the foundation for data mining. More generally data mining can be thought of as the process of creating predictive models from data. The holy grail of data mining is to produce very accurate predictive models

automatically on large data sets. We are not there yet, but we are certainly closer than we were just a few years ago.

2 Trend 1. More Models

What these means is that where as only a few years ago models were few and far between, we are now entering a world in which predictive models about customers and their transactions are becoming more common. This is the first trend: accurate predictive models are becoming more common, thanks in some part to the ability to produce such models using one of several statistical or data mining packages running on PCs. A decade or two ago, most statistical models were produced on mainframe computers, if only because that was were the data was.

With more models comes the need to integrate several models about a customer or transaction into a single score so that an appropriate action can be taken. For example, when a card customer logs onto a web site, the site has access to a variety of models about that customer. Usually, at most one or two actions can be taken, such as suggesting an offer or some content to view.

Sometimes this is called hierarchical modeling or profile based modeling. The goal is to integrate all the available data and produce a single best action. There are no vendors yet offering this type of modeling, but several should emerge over the next few years.

3 Trend 2. Better Models

Twenty years ago, the role of modelers was to understand the data well enough so that they could throw away 99% of it, ignore most of the attributes, divide it into five or six segments, build a linear model, guard the coefficients with their lives, argue to the marketing and risk VPs that this approach was cutting edge and created a huge competitive advantage, and tell their managers that they were underpaid.

Today, techniques such as tree-based classifiers and ensembles of models allow predictive models to built on larger data sets. In practice, this means that models can be built on transaction level data for the first time. These days good modelers understand the data well enough so that they need only throw away 70% of it, ignore fewer attributes, divide the data into dozens or more segments, build nonlinear models such as trees and neural nets, guard the coefficients with their lives, argue to the marketing and risk VPs that

this approach is cutting edge and creates a huge competitive advantage, and tell their managers that they are underpaid. This is progress. In general using more data to build models leads to more accurate models.

4 Trend 3. Less Expensive Models

Until very recently deploying predictive models was like doing business in certain countries in which the only way to make real progress is to spread bribes far and wide. This is simple to explain: predictive models are produced by statisticians using statistical software, while to be useful predictive models must be deployed in operational and back office systems by IT professionals using Cobol, and more recently C and C++. Since these two groups think differently and use different languages, projects to deploy and update models (usually described on yellow pads) got very low priority and essentially the only way to deploy a model in less than three to six months is to spread a lot of bribes and fear around.

Today, standards and middleware for predictive models are emerging. The Data Mining Group (www.dmg.org) is a consortium of vendors using a common XML format called the predictive model markup language (PMML) to describe common predictive models, such as logistic regression, tree based classifiers, and neural networks. This means that that models can be deployed independently of the software used to create them. PMML provides a platform, application and operating system independent way of describing predictive models.

More importantly, PMML separates the description of the model from the code executing it. What this means is that models described in PMML can be safely inserted in real time into 24x7 operational systems, something that is not advisable with alternative technologies such as COM, CORBA, or component based approaches.

Middleware is emerging for deploying PMML models in which one can integrate the middleware once and deploy the models immediately. PMML and the emerging deployment middleware should dramatically reduce the cost to deploy new models and update old models.

Acknowledgements

This is a 2002 revision of an article that first appeared in Card News in 1999. Some unauthorized variants of this article have appeared on the web in different forms without permission.

Copyright

This article is copyrighted by Robert L. Grossman, 1999, 2002.

About the Author

Robert Grossman is the Founder and a Partner of Open Data Group, which provides consulting services, outsourced data services, and litigation support services related to data. He is also the Director of Informatics at the Institute for Genomics and Systems Biology at the University of Chicago. He has written over 100 papers and edited four books in data mining, business intelligence, direct marketing, e-business, high performance computing, and related areas. He has a Ph.D. from Princeton and a A.B. from Harvard.

For More Information

For more information, please contact Open Data Group at info at opendata-group dot com.