

## **Data Detectives: Number Crunching Alone Won't Build You a Successful Predictive Model**

*By Jim Wheaton  
Principal, Wheaton Group*

*Original version of an article that appeared in the May 1998 issue of "Catalog Age"*

### **Introduction**

Often, I am frustrated when talking with prospects who are interested in building a predictive model. Many have listened to the promotional pitches of certain data mining software companies and are predisposed to the "push button" approach to modeling. After all, automation will eliminate the need to interact with quantitative analysts who ask difficult questions, speak in strange tongues, and charge high prices for a product that they claim takes quite some time to create.

It is my strong opinion that there never will be a substitute for an experienced human analyst. This article will explain why by focusing on a critical, up-front part of the predictive modeling process called Exploratory Data Analysis.

Consider that:

- A model is a simulation of reality;
- Reality cannot be simulated if it is not understood;
- Reality cannot be understood without human involvement.

Sooner or later, any automated modeling technique will cause trouble. This is because it is easy to recognize statistical patterns within the data. What is difficult, however, is to identify the subset of patterns that make business sense and are likely to hold up over time. The only way to differentiate "true" patterns from those caused by the vagaries of sampling and transitory environmental circumstances is to hire a seasoned Data Detective.

Before proceeding with the main body of this article, I will make a statement that is sure to be controversial within certain factions of the data mining community:

- If you want to get rich, develop a "push button" analytical tool.
- If you want to build great models, concentrate on becoming an experienced Data Detective. This is because the ultimate goal of any such project is to model the business, and only incidentally to build the model.

## A Definition of Terms

Before beginning, the following is some terminology:

- The dependent variable is the behavior that the model is attempting to predict. Common examples are response, sales, profit, and lifetime value.
- Independent variables are the historical factors that have been incorporated into the model to predict behavior. These are also known as “predictor variables,” or “predictors” for short.
- Potential predictors are all of the historical factors that are interrogated by the analyst as candidates for ultimate inclusion into the model.

## The “Art” of Exploratory Data Analysis

There are two things that we do in Exploratory Data Analysis:

- Technique-specific manipulation
- Capturing the underlying dynamics of the business being modeled

We will not focus on technique-specific manipulation in this article. Although somewhat of a controversial position, I contend that is that this is an almost-clerical task. In linear regression, for example, the potential predictors must be transformed into what is known as a linear relationship with the dependent variable. In other words, they must plot out as roughly straight lines against what the model is attempting to predict. This can be handled almost without thinking, through standard mathematical manipulations to “straighten” the potential predictors.

Instead, the true art of Exploratory Data Analysis lies with determining whether or not the relationship of the underlying “raw” potential predictors makes ongoing business sense. Only those potential predictors that pass this test should even be given the chance to end up in the final model.

This is an intellectual process that separates the good analysts from the bad. Its essence lies with capturing the underlying dynamics of the business being modeled, and involves the identification of four things:

- Errors in the data
- Outliers in the data
- Anomalies in the data
- Potential predictors that do not reside on the database

The balance of this article is a sequential discussion of each of these topics.

## Errors in the Data

The model will be a disaster if it is based on erroneous data. The first step of a thorough Exploratory Data Analysis is a careful examination of the analysis file to ensure that it contains nothing but accurate data. Consider the following example:

- For a large and well-known direct marketer, a customer model was built off an analysis file consisting of four mailings, representing the first drop of each major season.
- As a first step, the analysis file – consisting of responders and non-responders for each of the four mailings – was interrogated for basic reasonableness (mail quantity, response rate, dollars per piece mailed, etc.). It immediately was apparent that something was wrong. Additional investigation revealed that, when the analysis file had been created, the response information had been appended to the incorrect mailings. The spring responders had been appended to the summer mailings, the summer responders to the fall mailings, and so forth.

## Outliers in the Data

Outliers are observations that are valid but not typical. Any cataloger has a handful of extremely loyal – sometimes even fanatical – customers. When constructing a predictive model, it is important to capture typical customer behavior. By definition, the behavior of these extremely loyal customers is not typical.

Consider a predictive model with sales as the dependent variable. To simplify matters, we will assume that the analysis file is comprised of just one mailing, and that responders have an average order size of \$50. Consider a customer who ordered \$5,000 of merchandise. Perhaps this is a salesman who found one of the catalog's items to be perfectly suited for a holiday-season gift to his customers. So, he ordered 100 of them.

Clearly, this salesman's order is an outlier. It is not typical. Including it – untreated – into the analysis file would not be a good strategy. This is because our salesman would represent an observation that would have the same amount of power as 100 typical responders in determining the final makeup of the predictive model.

Confronted with such outliers, the seasoned Data Detective will adjust the data in some way. Perhaps the outliers should be dropped from the analysis file. Or, maybe they should be included in a "capped" form – say, every order above \$250 will be counted as just \$250. If a "capping" strategy is employed, the Data Detective must then decide on the optimal level of the cap. These are the types of decisions that comprise the "art" of Exploratory Data Analysis.

### **Anomalies in the Data**

For each potential predictor, it is critical to determine whether its relationship to the dependent variable makes ongoing business sense. Frequently, a potential predictor will have a relationship to the dependent variable that is statistically strong but unlikely to hold up over time. Consider a direct marketer in which an extremely strong, positive relationship was found between ownership of the private-label credit card and response:

- Despite this strong relationship, the variable was left out of the model. At the time of the mailings that comprised the analysis file, the credit card had just been introduced.
- Therefore, the small number of card owners generally were the company's most fervent buyers. This represents a phenomenon called self-selection.

However, by the time the model was to be put into production, the card ownership universe had expanded significantly. Many of the subsequent holders of the card were not nearly as enthusiastic in their devotion to the direct marketer. Therefore, the relationship of card ownership to response had significantly weakened, making it an unstable potential predictor.

### **Potential Predictors that Do Not Reside On the Database**

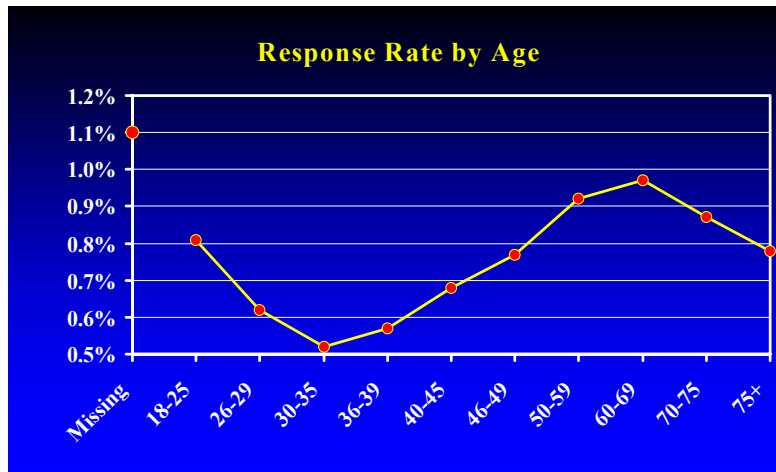
The "button pusher" will consider as potential predictors only those fields that exist on the database. The Data Detective will use his or her creativity to calculate additional predictors. Consider just a few of the potential predictors that can be derived from the Lifetime Number of Orders field:

- Breakouts by merchandise classes. (The determination of the appropriate classes is an "art" in its own right.)
- Breakouts by Kept versus Returned Orders. (The analyst then must determine how to handle orders in which some items were kept and others returned.)
- Breakouts by Phone versus Direct Mail
- Breakouts by Gift versus Non-Gift
- Seasonality measurements; that is, determining if some customers only order during the holidays

### **Putting it All Together**

The following example is not from the catalog industry. Instead, it is a prospect model that was built for the marketer of a credit card with tie-ins to hotel and airline discounts. Nevertheless, it is the best I have seen for illustrating the thought processes that are required for top-notch Exploratory Data

Analysis. I will outline the analysis that transpired for just one potential predictor. **The dependent variable was response. All of the potential predictors were derived from the demographic overlay data of a major compiler. For Age, the analyst created the following report of response rate by age:**



The following are the key characteristics displayed by this report:

- For 18 to 25 year olds, the response rate is relatively high.
- Then it declines, until bottoming out at 30 to 35.
- Beyond 35, response climbs steadily, finally peaking at 60 to 69.
- Beyond 69, it tails off.
- Response is at its highest among the prospects for which Age information is unavailable.

The analyst, a seasoned Data Detective, recognized that the relationship of Age to response was complex. Therefore, he was unwilling to include it as a potential predictor unless he was able to construct a cogent marketing explanation for the relationship. Subsequent analysis, involving insights from the client, provided just such an explanation:

- Young people are excellent candidates for credit cards offering travel discounts. They are not yet tied down by responsibilities such as kids, mortgages, and car payments.
- As they get into their mid- to late-twenties, they enter into the very demanding “Ozzie and Harriet” stage of life. Obligations abound, such as young children, substantial mortgages, and multiple car payments. It is an unlikely time of life to be traveling.
- From their mid-thirties to late-sixties, the opportunity to travel steadily increases. The kids are older – perhaps even out of the house. Also, family income is rising steadily.

- At about seventy, the desire to travel begins to wane. Health considerations play a major role. So, too, does the concern for dwindling financial resources.
- As for response being the highest among prospects with no Age information, this is a reflection of the un-Ozzie-and-Harriet phenomenon:

The extent to which an individual has children, a mortgage, credit cards, and the like, is the extent to which this individual is likely to be represented on a given demographic overlay database. Conversely, those individuals who cannot be matched to an overlay database tend to be young renters who move frequently. As we have already established, these individuals are very responsive to our credit card travel offer.

The analyst was satisfied by this marketing explanation. As a result, Age was allowed into the modeling process as a potential predictor. The variable was transformed mathematically so that it would display a linear relationship with response – essentially, a clerical task. Ultimately, Age survived to become an important predictor in the final model.

## Summary

There is no magical shortcut when building a predictive model. If you want good results, invest in the services of a seasoned Data Detective rather than a "button pusher." And concentrate on meticulous Exploratory Data Analysis.

I'll close with the following thought. Assume that, sometime in the future, data mining software is developed with unprecedented predictive powers. In this case, thoughtful Exploratory Data Analysis will be even more critical. Without it, our powerful data mining software will – by definition – do a superior job of identifying the spurious patterns that are inherent in bad data. Therefore, the resulting predictive model will point us even farther away from our true target market!

In other words, there is no substitute in the modeling business for an experienced Data Detective.

*Jim Wheaton is a Principal at Wheaton Group, and can be reached at 919-969-8859 or [jim.wheaton@wheatongroup.com](mailto:jim.wheaton@wheatongroup.com). The firm specializes in direct marketing consulting and data mining, data quality assessment and assurance, and the delivery of cost-effective data warehouses and marts. Jim is also a Co-Founder of Data University [www.datauniversity.org](http://www.datauniversity.org).*