

Myths and Realities of Building Models

*By Jim Wheaton
Principal, Wheaton Group*

Original Version of an article that appeared in the October 3, 1994 issue of “DM News”

The merits of various modeling techniques inspire heated debate. Recently, the proponents of regression versus neural network models have been slugging it out in the pages of various trade journals. It is time to put the rancor into perspective. Having participated in countless model builds, I believe that the specific technique plays only a secondary role in determining the success or failure of a model.

Whether the technique of choice is regression or neural nets (or, for that matter, tree analysis or fractal geometry), what really separates the good models from the bad is the up-front work that must be done before the formal modeling process. Also important is the backend process of correctly implementing the model, time and time again, in a production environment.

The following is a chronological listing of the steps required to build a predictive model, along with the estimated percentage of total project time:

Modeling Step	Pct of Time
Up Front Work	
1. Developing a Research Design	10%
2. Creating Analysis Files	30%
Model Build	
3. Exploratory Data Analysis	30%
4. Generating the Predictive Model (i.e., “Scoring Equation”)	10%
Implementation	
5. Deploying the Model	10%
6. Creating Ongoing Quality Control Procedures	10%

I have highlighted the 10 percent of the process that generates the scoring algorithm (i.e., the formal modeling process) because this is what the recent regression versus neural net battle is all about. But the fact is, if you short-change the other 90 percent you probably will end up with a lousy model.

Before I get into each of these modeling steps that comprise “the other 90 percent,” let me rephrase the previous paragraph in a way that I am sure will be controversial with certain factions of the modeling community. If you want to get famous, promote the latest statistical technique for generating a scoring algorithm. If you want to build a great model, concentrate on the remaining, unglamorous 90 percent of the process.

#1: Develop a Sound Research Design

Without a sound research design, you have nothing. And what is a sound research design? For starters, it is the identification of a solvable problem. It is remarkable how many companies look to a predictive model to solve a problem that no model could ever solve, regardless of the technique.

Consider, for example, one direct marketer's request to double the response rate of rental lists while maintaining outside mail quantities. In reality, this requires a marketing or merchandising revolution, not a model. It is quite possible to build a demographic/psychographic predictive model to find segments of rental lists that perform at twice the average. Unfortunately, these segments generally will not represent much more than ten percent of the outside mail quantity.

Having identified a solvable problem, the next challenge is to identify a subset of representative past mailings that will comprise the analysis file. As an extreme example, a fundraising mailing for a veteran's organization from the time of the Persian Gulf war probably would not be a good candidate for an analysis file. Response patterns during this atypical time in our nation's recent past are not likely to be typical. Other more common factors to consider include seasonality, creative formats, and merchandise mixes.

Another important task is to determine the optimal dependant variable. If available, sales is preferable to response. After all, responders are not equal in value. And, a net sales model is superior to a gross sales model. The one caveat is that it sometimes is not worth the wait for returns information before building the model. For example, a business with a returns rate of four percent is not likely to see much incremental benefit in modeling net versus gross sales.

Also, where multiple types of merchandise are offered with very different gross margins, net profit models are optimal. And finally, there exist circumstances in which cumulative long-term sales or profits should be modeled. This is particularly true of “contract businesses” such as subscriptions and continuities.

#2: Create Accurate Analysis Files

A perfect research design is worthless if the analysis file is inaccurate. And it is very easy to generate an inaccurate file. The process of appending response information to the mail history (“time 0”) file(s) often involves a number of complex steps that invite error. Also, the underlying database can be flawed. Although literally countless things can go wrong, a couple of “war stories” should provide context:

- A customer model was built off an analysis file consisting of four mailings. For each mailing, the analysis file was interrogated for basic reasonableness (mail quantity, response rate, dollars per piece mailed, etc.). It was immediately apparent that something was wrong. Additional investigation revealed that the response information had been appended to the incorrect mailings.
- For another customer model, interrogation of the analysis file revealed an unusually large percentage of individuals with only one order at the time of each mailing (single-buyers). The client's answer appeared reasonable: the business had enjoyed rapid and recent growth.

The real reason was discovered only later: about \$80 million dollars of transactions, representing about two and a half years of history, was unavailable when the database was constructed. The “live” model results suffered when who appeared to be the single-buyer inhabitants of the bottom deciles – who were in fact multi-buyers – ordered merchandise with a vengeance.

#3: Understand and Manipulate the Data

Even with a correct analysis file and a sound underlying database, the work involved in a model build has only just begun. Consider that:

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

This leads to the concept of automated modeling, which is a major focal point of the regression versus neural net debate. I am not biased against neural networks. In fact, Neodata's research and consulting group is currently testing one vendor's product. But I am biased against anyone who claims that neural networks are the key to automated modeling. These people argue that:

- Regression models require the transformation of each predictor variable into a linear relationship with the dependant variable. This is not necessary with neural nets.
- Regression models require additional steps to create interactions. This is not necessary with neural nets.
- Therefore, neural nets can be used in an automated fashion to create predictive models.

The truth of the matter is that the two aforementioned regression steps of variable transformation and interaction creation can be automated with relative ease. Therefore, a key perceived advantage of neural nets is eliminated.

But this is missing the point, which is that any automated modeling technique is likely to cause trouble. This is because it is easy to recognize patterns within the data. What is difficult is to identify those patterns that make sense and are likely to hold up over time. Consider the following two examples from a retail customer model:

Pattern #1: A weak negative relationship was found between response and customer distance from the nearest store.

Disposition: Good retail customers generally do not live far from a store. The distance variable represented the straight-line distance between each customer's ZIP Centroid and the nearest store's ZIP Centroid. It was theorized that, for many customers, this was not a sufficiently refined calculation. Distance was recalculated by Carrier Route Centroid, and the relationship to response went positive. Because this made intuitive sense, the Carrier Route version was used in the model.

Pattern #2: An unusually strong positive relationship was found between response and customer ownership of the in-store credit card.

Disposition: The variable was left out of the model. At the time of the analysis file mailings, the credit card had just been introduced. Therefore, the small number of card owners were generally the client's most fervent buyers. However, by the time the model was to be put into production, the card ownership universe had expanded significantly. Therefore, the relationship of card ownership to response would have changed dramatically.

#4: Transfer the Algorithm to Production

A model is worthless if it cannot be accurately transferred to production. One way to minimize this problem is for the analyst to work off the same record format as the production database. This eliminates the need to write translation code.

Unfortunately, many analysts do just the opposite because they find the format of the production database to be cumbersome. Often, their solution is to have a programmer create a fixed-length extract of the database record that contains only the fields required to build the model. "Recency," for example, might be in positions 25 to 27 of the extract but who-knows-where in the production database.

This is where the translation code comes in. Things get really interesting when the production database is in a sophisticated format such as variable-length/modular or relational.

While working off the same record format as the production database solves this problem, other difficulties may still arise. After all, database formats can change over time. Even more common is when the values within a given field change definition.

Consider, for example, the “department field” in a retail customer model. If the value of "06" corresponds to "sporting goods" on the analysis file but to "jewelry" on the production database, the model is not likely to be a success. This sort of change is all too common within the database world.

#5: Create Ongoing Quality Control (QC) Procedures

This is an extension of the previous step's discussion, because models remain in production for as long as two years. *[Subsequent, July 23, 2003, comment: I have seen models remain viable and in production for far longer than two years. Two years ago, for example, a former client rebuilt a model of fresh data to replace one built in 1994 off 1993 data. The two models validated identically!]* QC procedures have to be established to ensure that the production version of the model continues to function the way the analyst intended it to.

QC is an involved process that I cannot fully detail in this space. However, one of the most effective QC tools is to profile the model segments off the analysis file and compare them with profiles of the model segments in production. These model segments should retain consistent profiles for key elements over time. If this is not the case, there may have been changes within the database, subtle or otherwise, that require further detective work to unravel. Such inconsistencies between the analysis file and production model segments raise a red flag that can help identify potential difficulties before they become problematic.

The following is an extreme example of what can happen without QC procedures. A predictive model was built by a “10-to-1 Decile shop”; that is, by a research company that labeled its best buyers "Decile 10" and its worst buyers "Decile 1." Unfortunately, the rest of the direct marketing world rank orders from Decile 1 to Decile 10.

The model was forwarded to a service bureau that had no previous relationship with the research company, with written instructions to "pull off the top four Deciles.” The service bureau, mindful of industry standards, proceeded to select Deciles 1 to 4, which resulted in the worst 40% of the file being mailed!

With QC reports, this disaster would have been avoided. For example, the service bureau would have known that a problem existed when it saw the following profile of 'average time since most recent purchase:

Decile	Recency
1	53 months
2	44 months
3	30 months
4	25 months

Summary

There is no magical shortcut when building a predictive model. If you want good results, concentrate on the unglamorous basics: research design, analysis file creation, understanding and manipulating the data, and transferring the actual scoring algorithm into production with ongoing QC procedures.

I will close with a final thought: Assume that there does exist a statistical technique that is superior in recognizing underlying patterns within the data. In this case, the other, unglamorous 90 percent of the modeling process will be even more critical. This is because our hypothetical statistical technique will, by definition, do a superior job of identifying the spurious patterns that are inherent in bad data. Therefore, the resulting scoring algorithm will point us even farther away from our true target market!

Jim Wheaton is a Principal at Wheaton Group, and can be reached at 919-969-8859 or 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.