10 Key Tips for Building Successful Data Mining Solutions - Part 3

The following is the third in a series of articles on the top ten tips for building successful data mining solutions. The tips listed below are not in order of importance, but are listed in terms of how they might occur within a given data mining project. Our top ten tips are as follows:

1) Identify the Right Problem
2) Getting Stakeholders Onside
3) Creating Quick Wins
4) Understanding the Data
5) Judicious Use of Statistics
6) Combining Art and Science
7) Establishing Performance Benchmarks
8) Interpreting Results Correctly
9) Relative vs. Absolute
10) Action and Measure - Actioning the Solution

This article focuses on tips 7 - 8.

Tip 7: Establishing Performance Benchmarks

It is incumbent on the builder of any predictive analytics solution to demonstrate the performance of the tool/solution built. The more complex the solution, the more important it is to demonstrate its effectiveness. Performance needs to be demonstrated with tangible results that can be measured and understood. The Gains Chart, or decile table (see Table 1 below), has become the standard in evaluating the performance of analytics solutions:

Table 1

Gains Chart Report						
% of Validation Sample	# of Validation Records	Response Rate	% of Total Responders	Response Rate Lift	Interval ROI	Modeling Benefits ($)
0% - 10%	20,000	3.50%	23%	233	145%	$ 26,667
11% - 20%	40,000	3.00%	40%	200	75%	$ 40,000
21% - 30%	60,000	2.75%	55%	183	58%	$ 50,000
31% - 40%	80,000	2.50%	67%	167	22%	$ 53,333
41% - 50%	100,000	2.25%	75%	150	-13%	$ 50,000
-	-	-	-	-	-	-
-	-	-	-	-	-	-
-	-	-	-	-	-	-
-	-	-	-	-	-	-
91% - 100%	200,000	1.5%	100%	100	-58%	$ -

The Gains Chart is produced by scoring each record in a Validation (holdout sample) of prospects/customers randomly selected and not used in the development of a predictive model. The validation sample is similar in every way to the sample records used for model development. In the example above, a model was used to score the 200,000 customers in the validation sample. The records in the sample are shown rank-ordered by model score (from highest to lowest) and grouped into 10 equal size buckets. Decile 1 represents the highest scored records and decile 10 the lowest scored records. The Response Rate column represents the actual observed response rate of the records in a given decile that occurred in a previous marketing campaign. If a model is performing very well, there will be higher response rates in the lower deciles (e.g. 0 - 10%) and lower response rates in the higher deciles (e.g. 91 - 100%).

Rank ordering by response rate is the actual model deliverable. We can then incorporate business measures like “Revenue per order” and “cost per marketing effort” to create a measure of Return on Investment (ROI). Results in the Gains Chart demonstrate the score cut-off point for records that deliver higher than average ROI. In the example above, those records in the top 40% of scores generate a positive ROI, while those below the top 40% generate a negative ROI.

The model created becomes the new benchmark for performance. The model’s performance must be continually assessed to determine whether its power (predictive ability) is eroding. Newer approaches and techniques should be measured against the initial model to establish new performance benchmarks. This is done by looking at the rank ordering capability of the model when compared to other scenarios. This is illustrated in Figure 1:

Figure 1

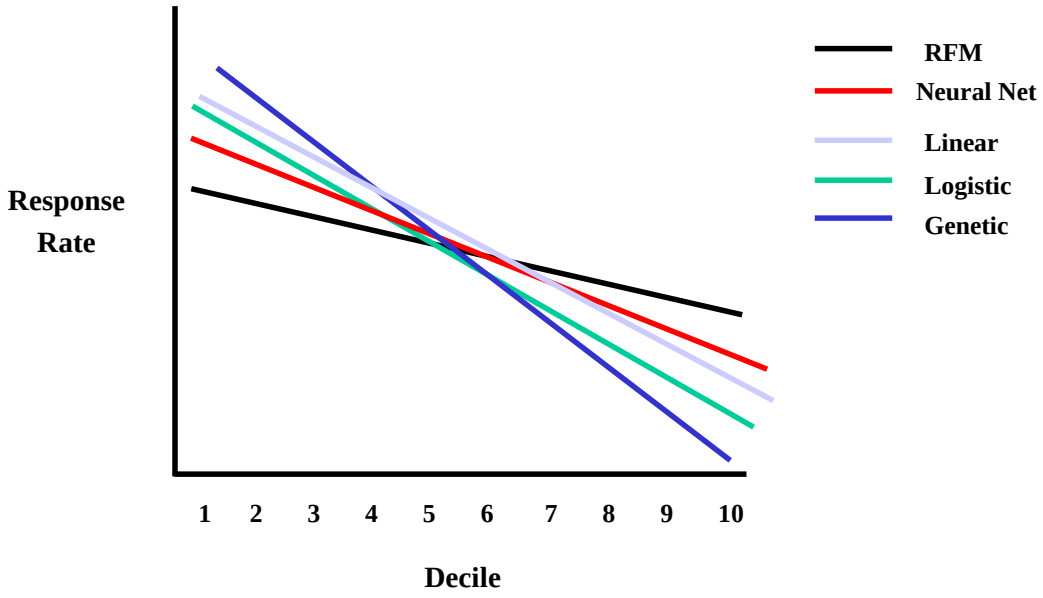

Figure 1 provides an example of the performance for five different response models derived from using five different techniques. A Lorenz curve or line, which plots the actual observed

response rate at each decile, is created for each model. A line with a descending (steeper) slope represents better rank ordering of the scored records and ultimately a better model. Looking at these above results, we would conclude that the best model for predicting response would be the one using the genetic algorithm approach.

It is critical to use metrics that are meaningful and tangible to the business when developing predictive tools and comparing the relative performance of them. Looking at the statistical diagnostics is important in building models, but it is the business metrics that are most easily understood that will determine the direction of predictive analytics solutions in a given business.

Tip 8: Interpreting Results Correctly

There is an often used, humorous phrase that is used to represent how statistics can be misleading -

"There are lies, more lies, and then statistics."

There is some truth to this saying as numbers can often be interpreted differently by different people. Numbers may point to a variety of problems with the data as well:

Table 2

% of Names Promoted	% of Mortgage Insurance Buyers
0 - 10%	80%
11 - 20%	5%
21 - 30%	3%
---	
90 - 100%	0%

In this mortgage insurance response model, we observe that the top 10% of names scored is capturing 80% of all the responders to this campaign. This kind of result raises all kinds of red flags. Typically, a high-performing model generates 25% to 35% of business within the top 10% of scores. A tool that generates **80%** of all the responders in that campaign in the top 10% of scores is suspicious. More probing and exploration of the data would be necessary in order to better understand why these results occurred. This might entail that the analyst actually scroll through the records and variable fields of the analytical file. Oftentimes a result like this is due to the presence of records that contain the dependant variable - previous Mortgage Insurance buyers. This problem can occur when the file created for analysis contains no pre-period that looks at behavior (Mortgage insurance purchase) prior to the response event (post period). Because of this failure to add a time dimension (pre and post periods) to the analytical file, all mortgage insurance responders would, appear as previous insurance buyers.

In this example, instead of providing valuable insights, the results led to questions of possible overstatement of performance. It is essential that someone with the appropriate skill set to 'actually get under the hood' of the data in order to identify what has caused this outcome and to resolve it accordingly. As practitioners, we pride ourselves in being able to explain a solution. This means that the actual variables or characteristics comprising a given solution should be easy to understand and their impact on the predicted behavior makes sense.

Correlation analysis reports and Exploratory Data Analysis (EDA) reports can easily demonstrate the actual impact of a given variable on the behavior in question. Multi-collinearity (an interaction between two variables) however, can cause variables within a model to switch in sign i.e. they have the opposite relationship with the target behavior when in a model, compared to their relationship with the target behavior when viewed alone (correlation) against the target behaviour. For example, different analyses (Correlations EDA's) may reveal that age and income are positively related to response to a given offer, implying that older people and higher income people are more likely to respond to the offer. However, when both variables become part of a multivariate solution, we may find that the actual sign for income is negative -implying that lower income people are more likely to respond. This switch in sign and relationship to the modeled behaviour needs to be understood in terms of what is causing this switch. Further investigation, including a correlation analysis between these variables (income and age), may reveal that there is a very high degree of correlation between age and income. When this occurs, two courses of action can be taken:

1. The first looks at replacing one of the variables (e.g. Income) with another variable that maintains the performance level of the model but does not change signs when combined with other variables that predict the modeled behavior. that is observed in our other analytical reports (EDA and correlation).
2. The second course of action would be to combine both age and income as one variable using techniques such as CHAID whereby the output of CHAID would determine how these variables should be combined.

In some cases, important characteristics may be overlooked when variables exhibit no statistical relationship in preliminary reports. One reason for this is the presence of "outliers" in the data.

Table 3

Response Variable	Spending (previous year)
Correlation Result	0.009
Confidence level	10%

Here the correlation results (Table 3 above) would indicate that Spending level has minimal impact on response, yet the corresponding Exploratory Data Analysis report (EDA) in Table 4 below would appear to illustrate a rather strong trend between spending and response:

Table 4

Exploratory Data Analysis Report	
Spending Level	**Response Rate**
$0 - $250	1%
$251 - $500	2%
$501 - $750	3%
$751 - $20,000	4%

The $20,000 spend level is clearly an extreme outlier that is diminishing impact of this variable in the data. Capping the spend value at $1,000 would resolve this problem and result in this variable having a stronger statistical relationship with response. A mix of automatic routines and some manual intervention used to arrive at an appropriate cap value. Ignoring outliers can result in the loss of key variables that produce a better predictive analytics solution.

The importance of this tip in interpreting results cannot be overemphasized. Solution builders within predictive analytics can sometimes have more of an academic focus that generates a pure result that can not be implemented. This is particularly the case if new analysts fresh out of school are building these solutions. The challenge for most organizations is to balance both the mathematical and experienced practitioners' viewpoint. Without this dual perspective, predictive analytics solutions in many cases will not achieve their intended effect.

Tip #8 represents the core of all data mining exercises. The benefits of any data mining exercise are only going to be as good as the data inputs. Without a clear understanding of what comprises these data inputs, incorrect insights and faulty conclusions may be drawn from the analysis and will yield decisions that may be sub-optimal, and in some cases detrimental to the organization. A disciplined approach to understanding of the data must be taken in this process so that the end result is the ability to conduct sound analytical exercises.