10 Key Tips for Building Successful Data Mining Solutions - Part 2

The following article is the second 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 4 - 6.

Tip 4: Understanding the Data

This tip 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 to understanding the data must be taken in this process so that the end result is the ability to conduct sound analytical exercises.

The process begins initially with loading the raw data source files. An initial review of the content of a source file can uncover some basic issues. For example, postal code values may contain data that is entirely numeric and not in the alpha numeric format that is the Canadian postal code standard. This kind of finding would suggest that the data may consist of non-Canadian customers or that there are wrong inputs in the postal code field.

In any data mining exercise, practitioners need to understand missing values, data formats and the number of unique values in each field. Boire Filler Group creates a Data Diagnostics Report comprised of data reporting tables to address these issues:

Table 1 - Data Diagnostics Report

Variable	# of records	Data Field Format	# of Unique Values	# of missing values
Income	100000	numeric	50000	2000
Customer Type	100000	character	4	10000
Gender	100000	character	2	50000
Household Size	100000	numeric	7	90000
Product Type	100000	character	3000	5000
Customer Name	100000	character	100000	0
Postal Code	100000	character	50000	0

In this report, the diagnostics are indicating that there are issues with household size and gender fields due to the large number of records (90% and 50% respectively) with missing values. One approach in dealing with this issue would be to not use these variables in any future analysis. Another approach would be to simply create binary variables for both fields where the variable merely looks at whether or not there is a reported value. In our experience, the creation of binary variables on the reported values can provide analytical insight. This is not that surprising if one considers that a person proactively reporting a particular value on an application or other type of report probably exhibits different behaviour than someone who chooses or neglects to report a value.

The other important finding relates to the column on the # of unique values for each variable. The key insight in interpreting these numbers is whether or not there is more than one unique value and if a character type variable contains a large number of values. Obviously, any variable with only one outcome or value is not going to be useful in any data mining exercise. A good example of this would be the use of gender in targeting insurance products to NHL hockey players. In the case of character type variables that have too many values, the creation of binary type "yes/no" variables for each value produces too *few* observations that actually have the 'yes' outcome. In order to provide meaning to the variable, there needs to be a way to combine these character values into meaningful broader groups. More about this topic will be discussed in the next tip called 'Judicious Use of Statistics'.

Another set of diagnostics reports are Frequency Distribution reports. These reports reveal how the values of a variable are distributed among the records in the file. The example below in Table 2 indicates that "Tenure" was not reported prior to 1998 and that product B seems to be the most prevalent product amongst this group of records.

Table 2 - Frequency Distribution Reports

Tenure	# of Customers	% of Customers
1998	9800	14%
1999	10000	14%
2000	12000	17%
2001	8000	11%
Missing	30000	43%
Total	69800	100%

Type of Product/Services Purchased	# of Customers	% of Customers
Product A	35000	29.66%
Product B	40000	33.90%
Product C	25000	21.19%
Product D	15000	12.71%
Other	3000	2.54%
Total	118000	100.00%

Another report called a “Database Cohort Report” - provides a view of the type of data that is updated an ongoing analysis. An example of data that is updated on an ongoing basis would be post campaign tracking type reports or standard KBM (Key Business Metric Reports) that are generated at specific intervals throughout the year. The intention of the database cohort report is to quickly observe if significant changes have occurred within the database over time.

Table 3 - Database Cohort Report

	Period 1	Period 2	Period 3
Record Count			
Average Purchase Amount			
Average Age			
Average Tenure			
Etc.			

Tip 5: Judicious Use Of Statistics

It is critical for both the data miner and business user in a project to understand the results generated by statistical insights. Blindly using statistics may generate learning that can not be actioned or worse, conclusions that lead to actions that are detrimental to the organization. For example, correlation reports (see Table 4 below) are designed to communicate the relative importance of a given variable (characteristic) against a desired behavior (e.g. response to a mailing). Variable correlations should provide a “profile” or description of the characteristics most closely associated with a “responder”. The correlation report in our example here would tell the following story about what a responder might look like:

- Has been a customer for a long time (Tenure)
- Tends to spend more
- Has bought a large number of products
- Is older
- Has Lower Income
- Credit Score has no impact on response behaviour (Credit Score not statistically significant)

Table 4 - Correlation Report

Variable Correlations			
Variable	Sign (relationship)	Correlation Co-efficient	Statistical Significance
Tenure	+	0.200	0.99
Spend	+	0.150	0.99
# of Products	+	0.130	0.98
Age	+	0.120	0.97
Income	-	-0.100	0.95
Credit Score	+	0.002	0.12

In building models with the above variables, the statistical significance of each variable and the interactions between these variables are important considerations in building the final model and the parameter estimates (weights) attached to each variable. For example, in the Final Model Variable Report (below) Tenure is by far is the strongest model variable accounting for 60% of the model's variability, and obviously has a much stronger relationship with response than indicated by the correlation results alone. So what is going on? Interaction effects between all the variables (commonly referred to as multi-collinearity), causes the model equation to result in only 3 variables, with Tenure accounting for a significant portion of the model's power.

Table 5 - Final Model Variable Report

Final Model Variables		
Model Variable	Impact on Response	% Contribution to Model
Tenure	+	60%
Spend	+	22%
Income	-	18%

Looking at both the Correlation Report and the Final Model Report, we can see that the modeling results convey a slightly different story than the correlation results, but this is acceptable as the results in each report will be used for different purposes. If we want to target specific prospects/customers, then we would want to use the model to obtain the right names. But if I want a more complete description of what a responder looks like, correlation results would be the more appropriate information to use.

As mentioned in Tip #4 - 'Understanding the Data', some fields or variables might contain numerous outcomes that are character values and not numeric. This would suggest that we create binary yes/no variables for each possible outcome. In some cases however, these variables may have too many yes/no outcomes, with very few records actually having a 'yes' value. There needs to be a way to group these outcomes into more meaningful categories. CHAID, a statistical routine that is often used to build predictive models, can be used in this context to create these categories.

Other statistical approaches for this purpose include “Product Sequencing” and “Affinity Analysis”. In these applications different types of correlation reports are produced to provide the necessary solutions. T hese type of analytical exercises are very data intensive and result in multiple iterations of correlation type reports. The iterations are produced to present different perspectives of the data. For example, we may want to look at product sequencing and affinity behaviour by value segment in order to see if there are changes in these types of behaviours across these segments.

Finally, “Advanced segmentation” employs the use of techniques such as clustering with the end result being the creation of distinct customer segments, whereby prospects/customers in each segment exhibits the same behaviours and characteristics. Any analyst building cluster solutions will tell you that there is both art and science in arriving at the right number of clusters as well as describing the characteristics of each cluster. This concept of using art and science actually leads us to our next tip.

Tip 6: Combining Art and Science

All experienced practitioners in this area will agree that there is blend of art and science in creating optimum solutions. Within each exercise or project, there are moments where the analyst looks beyond the science in arriving at approaches in building better solutions. The following are two examples.

The first example is the adoption of the old adage ‘Fish where the fish are’. To the practitioner, this might imply that a good place to commence any analytics is to explore those areas that have a known high level of customer penetration. More specifically within the area of new customer acquisition, exploring the notion of existing customer penetration as a means of targeting new customers certainly represents one sound analytical approach. Future acquisition efforts would focus on the selection of prospects that reside in geographic areas of high customer penetration. Of course, depending on the level of sophistication and data environment, other approaches can be used in conjunction with the ‘Fish where the fish are’ approach.

A second example is in the case where no customer data exists. Even the use of the ‘Fish where the fish are’ approach would not even apply here. So is there anything we can do here? Closer examination might reveal that market research information exists that indicates the purchasers of this company’s products have higher income, have a large number of persons in their household, and are immigrants.

The experienced analyst would realize that although customer data is not available, there is Statistics Canada information at the postal area level that relates to these characteristics. Using this knowledge, the analyst can create a composite index of these three characteristics in the Stats Can Index Report and target geographic areas based on those that index highest.

Table 7 - Stats Can Index Report

StatsCan Index Report			
Model Variable	**Income**	**% 3+ Household**	**% Landed Immigrant**
Average all Postal Codes	$ 40,000	52%	5%
Average for M5J 1A2	$ 50,000	60%	10%
Index	**1.25**	**1.15**	**2**

This composite index or score for a prospect residing in M5A1J2 would be (.33 x 1.25) + (.33 x 1.15) + (.33 x 2) = 1.45. Each prospect/customer would obtain a score using this method. This kind of indexing approach could be used to provide some means of targeting for acquisition type programs that essentially leverages market research learning, when no customer information is available. Although there is some leap of faith in assuming that market research behaviour compares similarly to existing customer behaviour, it represents a better approach than targeting at random, and certainly warrants testomg within an acquisition program.

From the examples above, it can be seen that intuition and judgment based on practical experience is a valuable component used to build an effective solution. The science component in a live campaign is key in evaluating the effectiveness of these 'art' driven approaches.

Please see our paper that provides a review of tips 7-10.