Myth or Fact: The Diminishing Marginal Returns of Variable Creation in Data Mining Solutions

Data Mining practitioners will tell you that much of the real value of their work is the ability to derive and create new variables from the source information within a database or file. For example, the calculation of averages or totals related to a specific timeframe or period represents information that is unlikely to be directly extracted from source files. Another good example is that of postal area variables that are based on 1st digit of the postal code. Although these examples seem pretty simplistic, it is not uncommon to have derived variables comprise over 90% of the information within an analytical exercise. From the data analyst's perspective, there are no limits to the number of derived variables that can be created. The limitation in variable creation is only confined to the imagination of the analyst or practitioner.

Obviously, creating variables and adding new information in theory should provide incremental benefit to any data mining solution, but as with any exercise or project, there are diminishing returns as one begins to explore the many possibilities and permutations that exist within the variable creation exercise. In this article, I will attempt to address this issue not in an academic manner but in the practical sense of whether it provides incremental business benefit to a given data mining solution. Specifically, I explore the impact of exploding the number of variables beyond our traditional techniques of variable creation.

The Traditional Techniques of Variable Creation

Our traditional techniques attempt to yield insights by creating variables in the following manner:

1. Binary Variables (yes/no outcome which represents the occurrence of some activity or event)
2. Average/Median or Sum of Variable
3. Index/Ordinal Variables whereby variable outcomes and values are placed into ranked groups. For example, age might be grouped into three outcomes with 1 being under 30 years, 2 being 31-50, and 3 being 51+.
4. Change/Velocity variables that look at how behaviour has changed overtime.

Across these four areas, it is not unusual to discover that the analyst has created hundreds of variables for a given analytical exercise. But are there other transformations that we should consider when building solutions? The notion of looking at a whole new suite of additional variables is that this potential new information can provide incremental benefit/insight to a given solution. In our research, we looked at predictive models that have been built using our traditional approach to variable creation note above.

Exploding the Number of Variables

We then looked at additional approaches that would explode the number of variables in our analytical file. Further mathematical transformations were employed consisting of the following:

- Log Transformation
- Square Root Transformation
- Sine transformation
- Cosine transformation
- Tangent Transformations

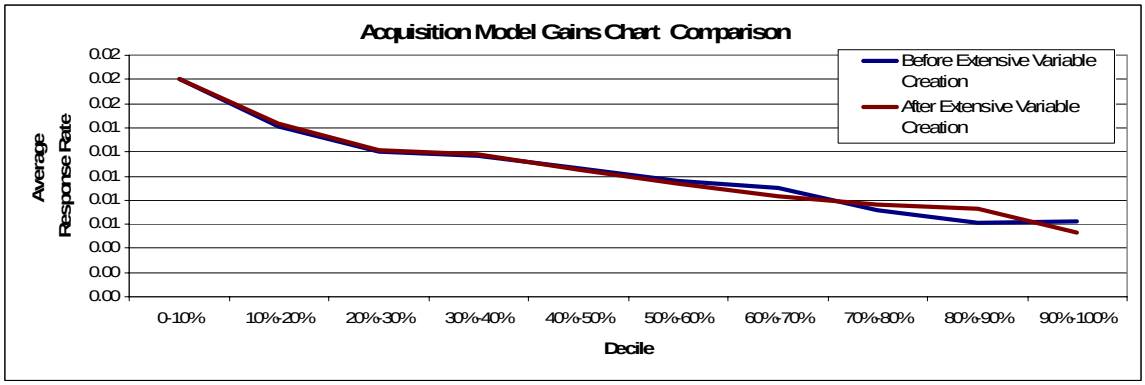
We also looked at combining pairs of top variables that were significant within the correlation analysis of the given predictive model. A good example of this is age and gender where we can actually capture the impact of age and gender together and observe their impact on the modeled behaviour. In the variable pair routines, we attempted to look at all possible combinations. If there are 20 variables that are derived using the traditional approach, then the potential number of possible variable pairs is 190 (20 X 19)/2.

Both these transformations (mathematical and pairs) dramatically increase the number of variables. The number of variables increased to 100 using the mathematical transformations (5 X20) and 190 (as seen above) for the possible number of variable pair transformations. In this simple exercise, 20 variables using the traditional approach now increases to 310 (20+100+190).

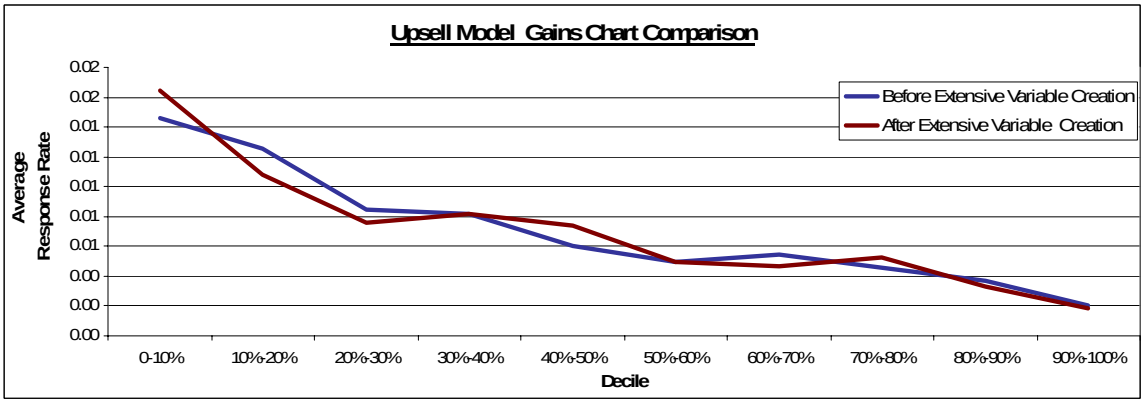
Business Cases to Demonstrate the Point

The challenge in this type of exercise is to determine if there is a real business benefit in exploding the number of variables from 20 to 310. The approach we took to assess this benefit was to look at models that were developed prior to this variable creation explosion and compare them to models that were developed after this variable explosion. Four models were looked at that were built by our company:

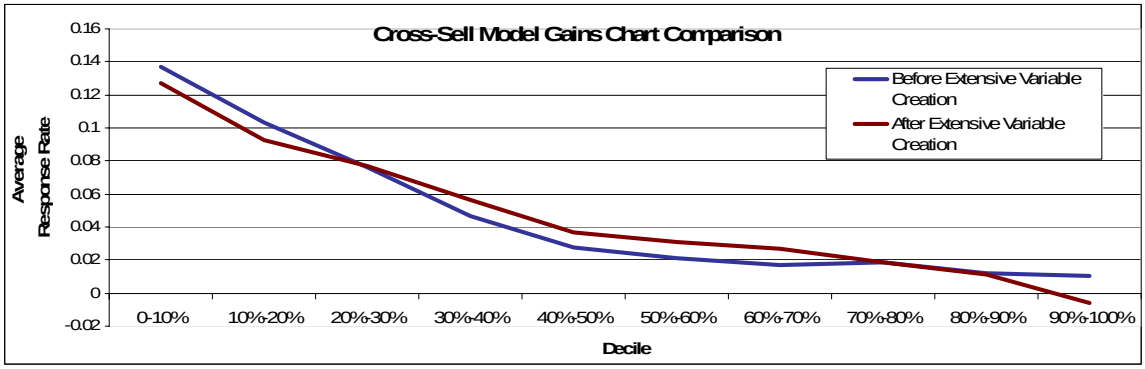
- Customer Acquisition Model



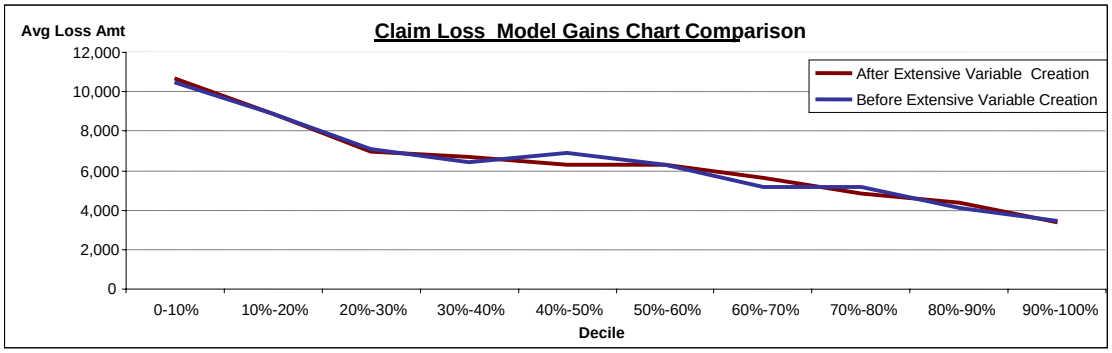
- Product Up-Sell Model



- Product Cross-Sell Model



- Insurance Claim Loss Model



The charts for each of these four models represent decile or gains chart reports where a sample of names is sorted by the model score into 10 deciles, with decile 1 representing the top (highest) model score and decile 10 the bottom (lowest) model score. The deciles are recorded on the x-axis and the observed model behaviour is plotted on the y-axis. As discussed in previous articles, a good model results when the line is quite steep and trending downward from decile 1 to decile 10. Our results from these four models indicate that that the lines or plots are virtually identical for models before and after the variable explosion. There is no tendency to have a more downward slope (better model providing more differentiation between customers) with the explosion in variables. This would indicate that there is no significant business benefit to increasing the number of variables using mathematical transformations or creating potential variable pairs.

But besides looking at the numbers and results, we may still want a more complete understanding of why these results occur. Let's look at variable creation in each of its stages.

Considering Demographic variables

The first stage actually represents the source information or variables that are not manipulated by the analyst. Good examples of these variables are, customer tenure, income, household size, etc. This information is extremely useful as it captures the basic demographics of the individual.

The subsequent stages all comprise derived variables. For now, we are going to focus on stages that deal with the following:

1. Grouping of variables
2. Calculating basic arithmetic diagnostics (mean, median, standard deviation, min and max)
3. Calculating change variables

1. Grouping of Variable Values

This stage looks at variables that represent the grouping of values into categories. A good example of this is postal code where postal codes are grouped into regions (i.e. all postal codes beginning with the 1st digit 'M' represent Toronto postal codes). Other examples might represent the grouping of specific product types or service codes into much broader product or service categories.

2. Calculating basic arithmetic diagnostics (total, mean, median, standard deviation, min and max)

This stage deals with the ability to summarize numerical information into a meaningful metric. However, it is only meaningful if we have historical information that can be used to calculate these kinds of diagnostics. For example, if we are calculating the average spend or the variation in spend, the question we must address is what timeframe are we looking at? Is the average or variation based on 6 months, 12 months, etc.?

3. Calculating change variables

Extending this logic of using historical information, we may want to identify how summarized behaviour changes over time. Has spending or product purchase behaviour changed over a period of time? Has it changed drastically over the last 3 months, over the last 6 months, etc.

Putting some perspective on the Traditional Approach to Variable Creation

In each of these stages, key information is being produced that is unique in explaining the behaviour we are modeling. Let's explore this thinking in more detail. The source-level information in many cases yields demographic information such as age or customer tenure that at one time or another have been key model variables. Grouping of values like assigning postal codes into regions, can produce more meaningful insight when attempting to look at geography in a broader context. Arithmetic diagnostics look at how summarized behaviour regarding key metrics can add value to a desired modeled behaviour. We all have seen examples of summarized metrics such as total products purchased or average spend as key variables within models. The diagnostic type variables differ from source variables in that source variables look at information at a point in time while diagnostics look at information over a period of time. Meanwhile, the change type variables add another dimension in that we are looking at how this summarized behaviour changes over time. Each of these stages, demographics, point in time variables from the source data, summarized data from calculating basic diagnostics, and change variables from the summarized data, represent unique ways of looking at the information. Because of these different unique perspectives, models will typically incorporate variables from each of these approaches.

Extensive Variable Creation

By providing some rationalization that the creation of variables using the traditional approach adds significant value to the modeling process, one may begin to ask whether a more extensive process can continue to add value to the process. Based on our analysis, it appears that this extensive expansion of variable creation provides minimal value to any model. The results are indicating that there is no real additional unique information that delivers incremental benefit to the model solution. It is our contention that *unique views* of information represent the real nuggets that add value to a data mining solution.

What does this mean going forward?

Considering that variable creation can be the most laborious part of any data mining exercise, these findings can provide some direction in how the analyst should best focus his or her time in any data mining project. Given the time pressure that analysts and practitioners face from business users, analysts can better focus their variable creation efforts and build solutions that are both timely and optimal.