

# **BEST PRACTICES IN MARKETING MODELING**

By

*Gregg Weldon, Chief Analytical Officer  
AnalyticsIQ, Inc.*



Over the last 15 years, there has been an explosive growth in the use of statistical modeling for predicting which consumers are most likely to respond and/or purchase products advertised by companies. These modeling techniques have progressed from a simple version of “judgmental modeling”, or assigning a random number of points for things that experienced marketers “knew” to be true, to today’s heavily-statistical output, generally built by well-trained statisticians who nevertheless may have little-to-no knowledge of the marketing space. Because the worlds of consumer marketing and statistical analysis don’t necessarily cross very often, it’s more important than ever for marketers to understand and review the work being performed by their modeling staff. This paper is designed as an overview for the marketing professional of the best practices in building statistically-accurate marketing models.

## **I. DATA ISSUES**

As in most endeavors, the most important portion of the modeling process is the preliminary set-up, or design, of the project. Is the goal of the model to predict which consumers will respond, apply, or convert? These are all very different results, and would result in completely different models. A sample design meeting should be set up before any analysis begins, so that all parties understand the goal of the particular project.

Associated with this is the question of what data is available for use. For example, let’s assume that the marketing firm is interested in building a model to predict which consumers will eventually convert (i.e.-actually purchase the product). For model development, the company would need a sample of converters (preferably from their most recent campaign(s)), a sample of responders who did not eventually convert, and a sample of mailed consumers who didn’t respond at all.

Who were these mailed consumers? Does the marketing company purchase list names that make up their universe of mailed consumers? Will these lists continue to be used in the future? How different will this campaign be from previous campaigns? Each of these questions needs to be answered at this early stage of the process. If the next campaign(s) will be much like the previous campaign(s), then building a stable model will be much easier. However, if there are major changes in products, list strategies, geographic mix, or a host of other items, the modeling process may be more problematic. These aren’t impossible to overcome, but they will need to be addressed fully so that all parties will understand what kind of results can be expected.

If this is the first time a campaign has been done (or it's a completely new type of product or area), so there is no previous data to model, the company may need to build a model on existing customers versus a random sample of the country (or at least a random sample within the company's geographic marketing area). In this case, any modeling results should be taken with a grain of salt, as it may take several months of experience to adjust any forecasts derived from the model to match actual conversion numbers.

To sum up, here are a few key questions that will need to be addressed:

**“What are we trying to predict—response, application, conversion, or something else?”**

**“How similar will our future campaigns be to our previous campaigns?”**

**“Do we have access to data for all of our status groups, including non-responders?”**

**“If this is our first campaign, can we access a random sample of non-customer consumers within our geographic footprint?”**

## **II. MODELING PROCEDURES – VARIABLE SELECTION**

Once the data issues have been determined, the modeler will be faced with figuring out which variables are most important in predicting the dependent variable (response, conversion, etc.). In most cases, there will be hundreds, if not thousands, of variables from which to choose. Narrowing down this list of variables will be key in moving the process along quickly and efficiently.

Data sources for variables may include demographic data, econometric data, credit data, or a whole host of others. In addition, there may be a mixture of household and/or area level data available. Some variables can be tossed out based on legal issues or logical concerns. For example, if building a model for an “invitation to apply” consumer offer, the use of individual credit data is not allowed legally. If building a model for a product offer for motor oil, having information about per capita use of lipstick for a geographic area may not be relevant.

Even with these obvious examples removed, the analyst will still have a huge number of variables in the dataset. The important element here is the determination of which variables are correlated with the dependent variable (what we're trying to predict) and which ones aren't. Bi-variate analysis, the practice of comparing a single potential variable with the dependent variable to see if they may be related, is the best way to narrow this list of variables.

### **Example: Bi-Variate Analysis**

Let's assume that the model being built will predict conversion versus non-response. A data observation for someone who did convert is given a value of “1” for the variable “Convert”. Non-responders are given a value of “0” for

Convert. Let's then take a look at the demographic variable "Age", which ranges from 18 to 99 years old. Bi-variate analysis will compare Convert with various ranges of Age to see if the age of a consumer has anything at all to do with if they will purchase the product being offered. Below is a breakdown of Age, once it's been formatted into groups:

<u>Age</u>	<u>% of Pop.</u>	<u>Avg. Convert</u>
18 to 24	18%	0.10
25 to 34	17%	0.14
35 to 44	10%	0.15
45 to 54	21%	0.20
55 to 64	20%	0.27
65+	14%	0.34

What we see here is that the older to person, the more likely they are to purchase this particular product. In fact, a 45 year old is twice as likely as a 24 year old to buy this product (0.10 vs. 0.20). In this case, bi-variate analysis states that, all other things being equal, Age is important in predicting conversion for this product.

Many times, bi-variate analysis will show the following. Let's assume that the variable in question is annual income.

<u>Income</u>	<u>% of Pop.</u>	<u>Avg. Convert</u>
\$0-\$20k	5%	0.18
\$20k-\$40k	20%	0.20
\$40k-\$60k	20%	0.19
\$60k-\$80k	35%	0.21
\$80k+	20%	0.19

In this example, annual income doesn't discriminate well between who's going to convert and who will not. Income isn't a factor in people's decision of whether to purchase the product at all.

Bi-variate analysis looks at the dependent variable (Convert, in this example) and the potential independent variable *in a vacuum*. This gives an indication that the independent variable may be important, but it doesn't tell us if it's more important than other variables we've identified through the bi-variate analysis process. This will be determined through multivariate analysis, during the regression phase of the modeling process.

### III. HANDLING MISSING VALUES

In a perfect world (or an academic setting), every variable in a dataset for every observation would be populated with a correct value. In reality, this is rarely the case. Missing values for variables is inevitable, and how these missings are dealt with can greatly affect a model's integrity.

Early in the modeling process, frequencies on variables are run to determine how many observations have values and how many are blank (or otherwise denoted as "missing"). Why is some data missing? This is the primary question that must be answered before proceeding. In many cases, a relatively small percentage of observations are missing a few variables. Often, they are missing in a "random" manner. In other words, there's no pattern to which observations are missing. In other cases, a few variables are missing a large percentage of the time. What happened in that case? The analyst must act as detective to track this information down.

For the following examples, assume a dataset with 10,000 observations (consumers). Of these, 5000 are converters and 5000 are non-responders (50% converter rate). Also, the variables to be used for modeling are 300 demographic variables taken from a variety of sources.

#### **Example 1: 5% Missing Each Time**

A frequency on a handful of the available variables reveals that 5% of the observations (500) are missing for each variable. These 500 consumers don't have values for any of the variables that will be used for modeling.

Why are these people missing each time? Do they have anything in common, such as geography? Are they more likely to be a converter or a non-responder? Is this simply a random event?

In many cases, it may be found that these observations are "just missing". There's no rhyme or reason for it. They're equally missing for converters and non-responders, and it's a relatively small percentage of the population. Since these people don't skew in any way, it's appropriate in many cases to simply delete these observations from the dataset and continue without them.

#### **Example 2: 5% to 10% Missing**

In this case, roughly 5% to 10% of the data is missing for almost all of the 300 available records. Analysis shows that no observations are missing on any regular basis; in fact, all 10,000 observations have appropriate values for most variables. However, no observation has an appropriate value for all 300 variables.

This is a typical situation with missing values. Data can often be missing in this way due to faulty collection, incomplete information, and common mistakes.

Because all observations have appropriate values for most variables, and no variable is missing more than 10% of the time, this should not be a problem for modeling.

### **Example 3: A Few Variables are Missing A LOT**

This is the point where “rules-of-thumb” get added to the mix for modeling. How much is “too much” when talking about missing values? Each experienced analyst has a different opinion on this. A good rule-of-thumb is 30%. By this, we’re saying that if a variable is missing just under one third of the time, the variable may not be the best one to use in a model. 30%+ missing is fairly high.

If your data, for the most part, looks like Example 2 above, but a few variables are missing over 30% of the time, then these variables need to be examined for possible exclusion from the modeling process.

Again, it comes down to finding out why these variables are missing so often. If the variable in question is “Monthly Mortgage Payment”, check to see if a missing value is equivalent to a value of \$0. Sometimes, data sources code 0 and (blank) interchangeably. Keeping the variable and treating missings as a value of \$0 is correct.

Or, these people may be renters, so the variable is null for them. In this situation, you may elect to treat missings as a completely separate category in the eventual model.

Thirdly, it may be that Monthly Mortgage Payment is always missing for non-responders and always present for converters. This means that the variable cannot be used for modeling, as the data is obviously skewed toward converters, not because converters have more mortgages than non-responders, but because the data for non-responders was in error and incomplete. This variable should be eliminated immediately from further analysis.

Three possible alternatives, three distinctly different ways to handle the missings.

### **Example 4: Almost All Variables Missing 30-35% of the Time**

Well, there are datasets where, even with a high percentage of missing values, you have no choice but to work with what you have. When almost all the variables are missing 30-40% of the time, the rule-of-thumb for how much is “too much” must increase. Now, you’ll need to look for whichever variables are missing significantly more than the rest. Suddenly, you’ve set your cut-off at 45% rather than 30% as before. The same analysis as listed in previous examples should apply, however. The goal of identifying any problems that may exist among missing variables is still in place. In fact, it becomes even more important to make sure that the variables you have are as correct and meaningful as possible.

#### IV. EDAs (CROSSTABS)

As noted in Part II, bi-variate analysis is very important in identifying variables that may be useful in predicting which consumers are more likely to become converters in the next marketing campaign. Analyzing each of these variables is called Exploratory Data Analysis, or reading “Crosstabs”. Crosstabs can be as simple as printing out a frequency of the independent variable (Age and Income were our examples above) crossed by the dependent variable (Convert=0 or Convert=1). When using any kind of statistical software (SAS is the industry standard), it’s quite easy to program crosstab output that gives much more information, which allows the analyst to make quicker, more educated decisions.

Let’s re-examine the example using Age.

<u>Age</u>	<u>% of Pop.</u>	<u>Avg. Convert</u>
18 to 24	18%	0.10
25 to 34	17%	0.14
35 to 44	10%	0.15
45 to 54	21%	0.20
55 to 64	20%	0.27
65+	14%	0.34

**Average** **0.21**

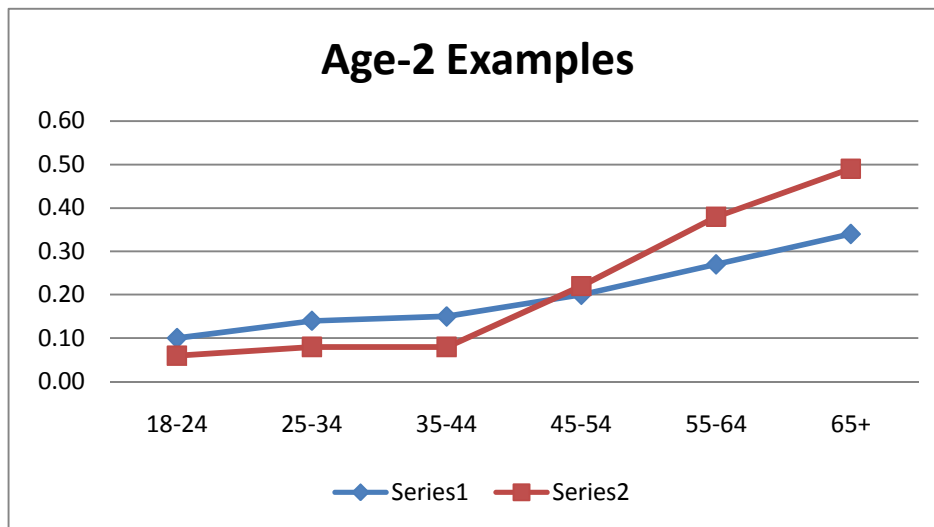
Age looks like a very good indicator of whether a consumer will become a converter or not. In fact, the trend is fairly smooth, as the older the consumer, the more likely they are to convert.

There are two ways to handle this variable. We may choose to use Age in a model on its own. That is, use it as a “continuous variable”, where the regression program would calculate points that would be multiplied by the actual age of the consumer. The older the consumer, the more points they would receive for the variable Age, reflecting the greater likelihood of their converting. The regression program will attempt to fit a straight line as close to actual Age as possible. Any difference between the actual line and the regressed, or estimated, line, makes up the error term. The goal of any model is to minimize this error term as much as possible.

The second method, using “Dummy variables”, breaks the variable Age into smaller subset variables that are then used in the model as blocks of data. For example, since the average conversion rate is 0.21, any group with a conversion ratio of less than 0.21 would be given negative points (less likely to become a converter) and any group with a conversion rate greater than 0.21 would be expected to get positive points (more likely to convert). For the Age example

above, we could conceivably create 4 dummy variables. Age1 (18 <= Age <= 24); Age2 (25 <= Age <= 44); Age3 (55 <= Age <= 64); Age4 (Age >= 65); Notice that we skipped the group of 45 to 54 year olds. These people have a conversion rate right at the mean value. They're considered our neutral group. By leaving them out of the model, they, by definition, are assigned 0 points. They're not more likely than average to convert (positive points) or less likely than average to convert (negative points). They're just average.

Dummy variables are used whenever the variable under consideration doesn't have a smooth trend like Age does in the above example. In many cases, a variable may have a sharp, non-linear curve, where the use of a continuous variable may not "fit" the line as well as a series of dummy variables.



Series 1 is the Age example we've previous reviewed. There is a fairly smooth trend, as older consumers convert more than younger consumers. Using Age as a continuous variable may be the best choice, as the statistical model will attempt to fit a sloped line that mirrors the actual Age results well.

Series 2 is very different, as people from 18 to 44 behave very similarly. Conversion rates really take off after age 44, though. In this example, the use of dummy variables would be more appropriate. A continuous variable would not be able to replicate the sudden lift-off at 45 that the data represents, leading to a higher error term and a less optimal score. Dummy variables, however, could be used to create subset Age variables. These would attempt to model a line from 18-44 (flat) and then a line for 45 or older (decidedly not flat). The use of these dummy variables would better minimize the error term by fitting 2 lines instead of one, leading to a more accurate score in the end.

## V. REGRESSIONS

Once the candidate variables for inclusion into the model are determined, and continuous and/or dummy variables are prepared, the analyst is ready for multivariate analysis. Multivariate analysis is simply the act of seeing how each independent variable affects the dependent variable (“Convert”) *as well* as all the other independent variables in the model. You may remember that bi-variate analysis looked at the relationship of each independent variable with the dependent variable separately and in a vacuum. Now, that vacuum is removed.

At this point, correlation between the various independent variables comes into play. When looking at crosstabs, it may have appeared that “Age”, “Average Age in Household”, and “Median Age at Block Level” all showed significant trends when measured against Convert. In multivariate analysis, however, these three highly correlated variables are all trying to tell the same thing about converters. This generally results in “white noise” that tends to actually harm the regression model being developed. Illogical results often occur as well, as the model may show negative points for older consumers and positive points for younger consumers, the exact opposite from what we know to be true from bi-variate analysis. This is because these variables are all trying to give us the same piece of the puzzle as to who is more likely to be a converter. A well-built, stable model will instead focus on a wide range of variables that each give a different and unique piece to the jigsaw puzzle that makes up the profile of a converter.

Once again, a “rule-of-thumb” enters the equation. A good one is that if two independent variables are correlated at 50% or more, once should be dropped from the model, leaving the other to represent that unique piece of the puzzle that the two had previously been sharing.

There are many types of regression available, but the two primary ones are linear regression and logistic regression. **Linear regression** is the most basic method, and it attempts to fit a straight line that comes as close to the data points in a sample as possible. Once this line is calculated, the distance between this estimated line and each actual data point is summed up to create the error term. The standard formula for linear regression is as follows:

$$y = a + bx + e,$$

where  $y$  is the predicted dependent variable,  $x$  is the independent variable,  $a$  is the  $y$ -intercept, or constant,  $b$  is the slope of the line, and  $e$  is the error term. In a typical model, there will be many more than just one independent variable, so there would be many more versions of  $(bx)$  in the formula.

Linear regression is typically taught in schools, as the basic formula is simple to understand and can be calculated by hand. Scores may range broadly, the higher the score, the more likely it is that the event will occur. On one model, a

score of 700 may be quite high and indicate the observation has a high likelihood of converting. On a different model, however, a score of 700 may be very low. The only way to see if a score is “good” or “bad” is to find where it stands relative to all other scored observations.

**Logistic Regression**, on the other hand, creates a score that is the probability of an event occurring. These scores range from 0 to 1, the higher the score, the more likely the event is to occur. Because the logistic score is a probability, a score of 0.80 means the same thing on any logistic scorecard; namely, that this observation has an 80% likelihood of converting. Logistic regression is very handy when building multiple models for the same client.

Performance-wise, linear and logistic regressions are identical. Neither method is “better” than the other. Linear is easier to grasp early on while logistic allows for multiple models to be compared directly.

## **VI. REGRESSION OUTPUT**

Whichever method is chosen, it is highly likely that the first regression run on the data will have variables in it that make no sense at all. It’s imperative that the regression output be compared to the crosstabs to see which variables “work” and which variables don’t.

The crosstabs, which were simply the bi-variate analysis of how an independent variable relates to the dependent variable, is the “answer key” for judging the model. If, for example, Age is in the regression as a continuous variable and, instead of giving a positive point value (which would reflect that the older a consumer was, the more likely they were to convert), we got a negative value, then Age is incorrect in the model. The trend in the crosstabs is the correct trend. The trend in the model **MUST** conform to the crosstabs, or be eliminated from the model entirely.

Why could Age, which had such a nice, strong trend in the crosstabs, look backward in the model? Correlation with other age variables, as noted above, is one major reason. It may also be the case that Age isn’t nearly as important as some other types of variables that are included in the model, so the value of Age as a variable is minimal. It all goes back to finding and fitting in all the pieces of the puzzle that makes up who is a converter. At the modeling stage, it becomes apparent which pieces of the puzzle are superfluous, which pieces are merely poor copies of other, more important pieces, and which pieces are “the real deal” in discriminating between converters and non-responders.

With each regression output, variables that don’t work (wrong signs representing incorrect trends, insignificant points) should be dropped from the model. Eventually, a model will be developed where each variable, be it continuous or

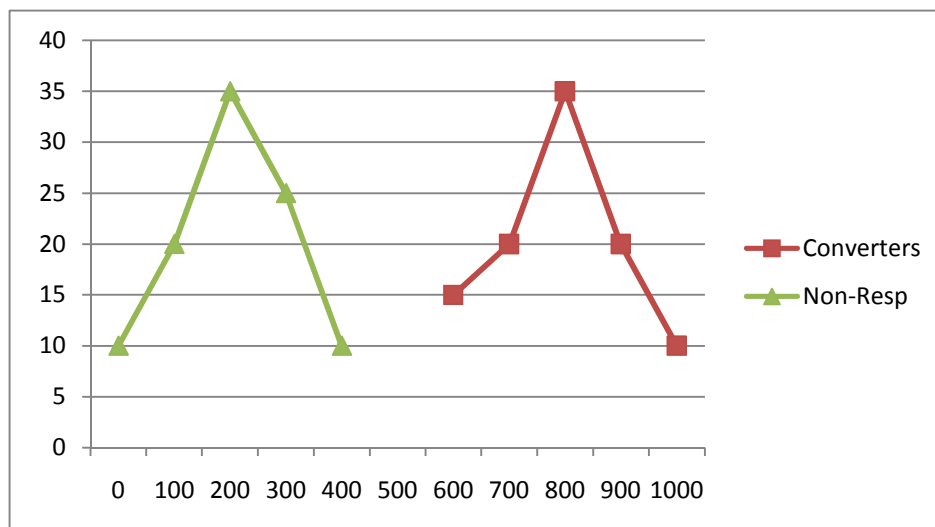
dummy, is statistically significant, has the correct sign, and is giving a unique picture of which consumers are most likely to become a converter.

## VII. MEASURING PERFORMANCE

Once a model is built, the question of how well the model actually works is sure to be brought up (by the client, at the very least!). Measuring the performance of the scorecard is very important. A good practice, early in the modeling process, is to split the data into two groups, a development dataset and a validation dataset. All work, from crosstabs through final regression, should be performed on the development dataset only. The validation dataset should be kept to the side and not brought out at all until after the model has been built.

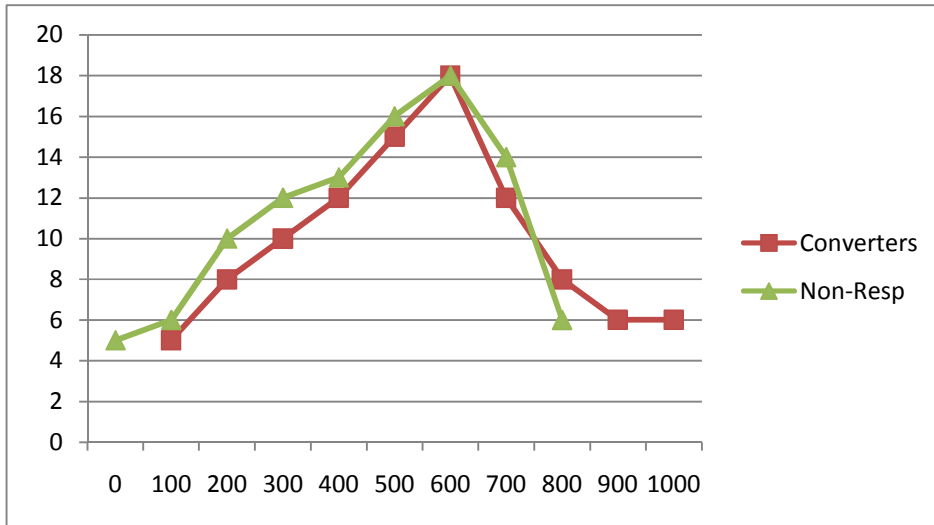
At this point, scoring the model on the validation dataset and seeing how it performs will indicate how well the model will actually work on independent data (i.e.- data it wasn't built on). There are several methods of measuring performance (KS, divergence, area-under-the curve, etc.). The KS test is most commonly used in marketing models. KS is easy to calculate, easy to explain, and easy for clients to understand. The KS score measures how differently the "goods" (converters in our case) and the "bads" (non-responders) score on the regression model.

Below are some graphs to help explain the KS score. The examples will assume that logistic regression was used in the model. Although logistic regression gives scores ranging from 0 to 1, industry practice is to multiply this score by 1000, so scores range from 0 to 1000.



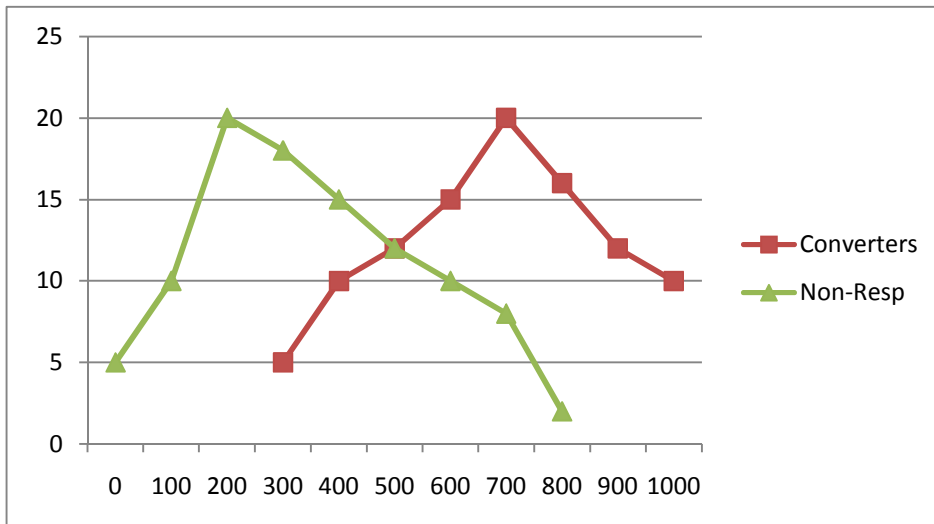
KS Graph 1

In Graph 1, we see a score laid out by converters and non-responders, with the score along the horizontal axis and percentage of the population along the vertical. In this example, all the non-responders (“bads”) scored under 500 and all converters scored over 500. A client would simply set a cut-off at 500, mail to everyone above that, and have a 100% conversion rate! This will never happen, unfortunately, unless a BIG error has occurred.



KS Graph 2

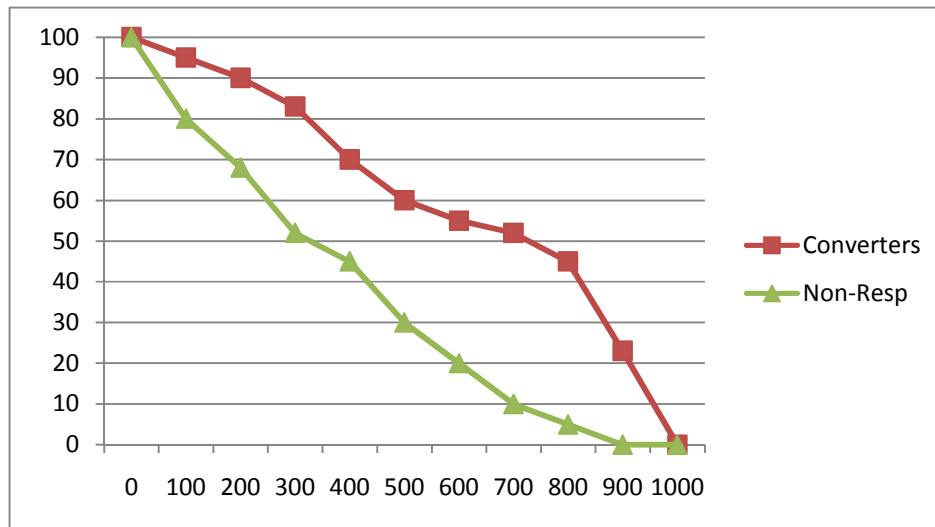
KS Graph 2 shows the opposite case, where there is no difference between the converters and non-responders. The model that created this graph doesn't work at all, as it shows that a client might as well as flip a coin rather than use this scorecard. Sadly, this situation DOES occasionally happen!



KS Graph 3

Most of the time, a model will look more like KS Graph 3. The non-responders as a group score lower than the converters, but there is overlap on both sides (Type I and Type II errors). The goal of the model is to separate these two groups as far apart as possible (like KS Group 1). The measurement of how far apart these graphs are is the KS Score.

KS is merely the measurement of the cumulative percentage of Goods (converters) minus the cumulative percentage of Bads (non-responders) at scoring intervals from 0 to 1000.



KS Graph 4

KS Graph 4 looks at the cumulative percentage of converters and non-responders at 100 point intervals of the score. We see that 100% of converters score above 0, 95% of converters score above 100, 90% of converters score above 200, etc., until we get to 0% of converters score above 1000, since 1000 is the maximum score. The graph also shows that non-responders go from 100% to 0% much quicker, indicating that they score lower on average than converters. For example, 100% of non-responders score above 0, but only 80% of non-responders score above 100, and so on.

The goal of the model is to create as wide a gap between these two lines as possible. The widest gap in this graph is at a score of 700. 52% of converters score higher than 700, but only 10% of non-responders score 700 or more, giving us a difference of 42. This value of 42 is called our KS Score. The farther apart those groups are pushed, the higher the KS Score, and the better the model is at discriminating between converters and non-responders.

If we had to measure the KS score on the first example (KS Graph 1), we'd have a KS of 100, as there is absolutely no overlap between the converters and non-responders at all. KS Graph 2 would have a KS approaching zero, as the lines overlap almost completely. KS Graph 3 is typically what you'd see in the real

world, and would look very similar to the cumulative results as presented in KS Graph 4.

## **VIII. CONCLUSION**

This article is intended to give the reader a summary of the best practices used in marketing models today. The subjects discussed should give a basic understanding of some of the difficulties and challenges involved when taking a mass of seemingly random data and creating from it something unique, logical, and profitable. Those interested in a more in-depth discussion on any of the topics touched on here are encouraged to further their research. Statistical modeling can be a fascinating subject, with applications in most industries today.

*Gregg Weldon is the Chief Analytical Officer of AnalyticsIQ, Inc. He can be reached at [greggw@analytics-iq.com](mailto:greggw@analytics-iq.com).*