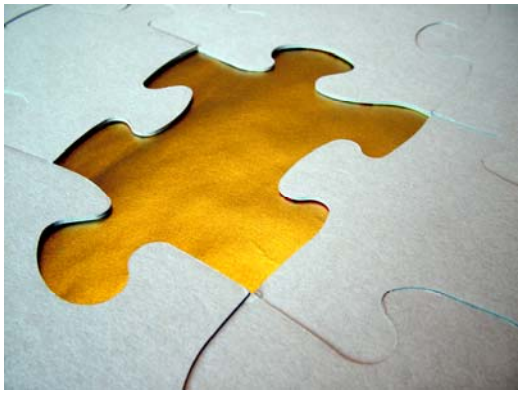
Missing Inaction … Can Be Hazardous to Your Data Mining Project

By Tim Graettinger



Did anyone ever tell you that you have "dirty data"? Were you offended? Surprised? Did you vow to do something about it? Or did you just resign yourself to living with it?

Missing data is one common component that contributes to the unpleasant moniker, dirty data. In this installment of my series on the nuts and bolts of data mining, we'll tackle missing data. We'll start first with detecting it. From there, we'll diagnose the issues, both qualitatively and quantitatively. With a proper diagnosis, we can then prescribe a treatment for any of the variety of situations that crop up. You'll walk away with a mental framework and a set of tools and techniques that are invaluable for real-world data mining applications.

What is Missing Data - and Why Do We Care?

Simply put, missing data or missing values, are "gaps" in a column of data. These gaps may be legitimate values, or not. To be more concrete, consider the excerpt of a data file shown in Table 1. In this excerpt, each row represents a health insurance policy holder. The various columns are attributes associated with the policy holders. The columns include:

REINSTATE	EFT	GENDER	PREM_AMT
	Y		$125
Y	Y	M	$300
	Y	F	$280
	N	M	
	Y		$170
	Y		$800
Y	Y	F	

Table 1 - Health insurance data excerpt

- REINSTATE –has the policy ever lapsed and been reinstated?
- EFT – is electronic funds transfer being used?
- GENDER – gender of the policy holder
- PREM_AMT – dollar amount of the monthly premium

In this case, the goal of the data mining effort is to build a retention model that predicts whether a policy holder is likely to stay or leave.

Notice that the REINSTATE, GENDER, and PREM_AMT fields in Table 1 have some null, blank, or otherwise missing values. Should you care? My primary goal in writing this article is to convince you that you should. Let the persuasion begin!

First, missing values can represent important information. In surveys, for instance, leaving a blank response to a question about income might actually signal a high-income individual. Or, consider a list of insurance prospects that has been appended with third-party demographic information. Some prospects in the list are not present in the third-party database, and hence, all the demographic information is missing for them. But, these people **are** often responsive to marketing efforts – due to the very fact that they **are NOT** represented on these large third-party databases (and, hence, are not receiving a lot of marketing pieces).

Second, some data mining methods (like logistic regression or neural networks) that expect numerical input data will barf and produce an error or warning when given a record that contains any missing data. Or, the method might "handle it" by simply ignoring the entire record – which may not be the desired behavior.

Finally, you **will** encounter missing values in your work. Two of my own recent projects are illustrative. In an insurance application, within a data set comprised of over 200 columns, 75% had missing values. A much different biotech modeling project was based on a more modest database having 40 columns. 100% of these columns contained missing values. In fact, "missing" was the most common value for every one of those columns.

Ignoring or mishandling missing values can have a powerful impact on results and performance. That's why you really need to know how to manage them.

Instruments of Detection

To manage missing values in your data, you first need to know where and to what degree they exist. That is, you need to detect the missing values. My approach is to do a "surface scan" of

the data. For me, that means running simple, standard statistics (mean, median, standard deviation, etc.) on all the numeric and date columns. Virtually all software packages will report the number of missing/invalid entries in each column.

For the non-numeric columns, I generally employ a software procedure to count the occurrences of each unique value in the field. Here, the results can require a bit more work to interpret because missing values in non-numeric fields are sometimes indicated in multiple ways, even within the same data set. For instance, missing might be represented as NULL (no character present in the data cell), or as a blank character, or as some other character or punctuation mark[1].

Finding the non-numeric columns with missing values can be a bit of an Easter egg hunt, but it is necessary and usually not too onerous. With the columns of interest identified, you can move on to diagnose them.

Diagnosis: Missing

What does it mean to diagnose the columns that contain missing data? And, why bother? To the first question, to diagnose means to understand (quantitatively) the degree to which missing values exist, and then to understand (qualitatively) why they exist. With respect to the "why bother?" question, we must diagnose before we prescribe. If we don't understand the missing value issues, we can hardly hope to apply the best treatment, can we?

Quantitative understanding of missing values starts by asking these two questions:

- How many columns of each type (numeric and non-numeric) have missing values?
- In those columns with missing values, what percentage is missing?

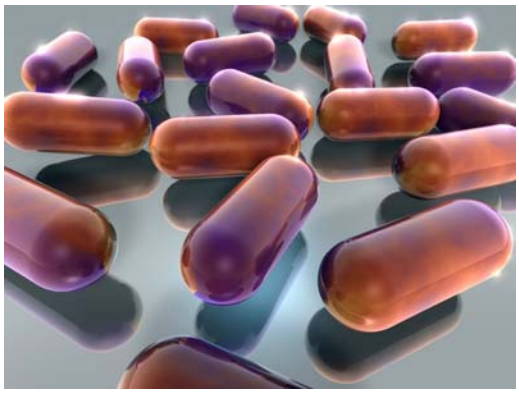
The answer to the first of these questions tells us how many treatments we will need to apply. Generally speaking, the more treatments, the longer it will take to complete them. The answers to the second question will point us in the direction of different treatments that will be described shortly.

On the qualitative side, we need to ask these questions (often together) of each column containing missing data:

- Why is the data missing?
- What does "missing" mean?

For instance, consider the GENDER column from the excerpt in Table 1. In asking "why", we learn that the missing values result from not matching the individual with a third-party list. For the PREM_AMT column in the Table 1 excerpt, by asking "why" about the missing values, we learn that terminated policy holders have had their values wiped clean since they are no longer paying premiums. That's a critical piece of information, and it represents a substantive, qualitative difference between rows (policy holders) in the data set. Without asking why, you might miss that crucial distinction – with disastrous results.

Prescription for Better Results



By asking the questions listed in the previous section, you are much more likely to take appropriate steps to handle missing values in your data. Let's take a look now at the workhorse options and the associated decision criteria for high-performance missing value management.

1. <u>When missing values are few and far between ... Do nothing</u>. Suppose that only a handful of columns have any missing values, and they occur only a miniscule number of times within those columns. In this case, most analysis and modeling procedures will simply ignore the missing entries. Since they are very infrequent, they will have no significant impact on the results.

2. <u>When a column has a significant number of missing values ... Create a missing/present (0/1) indicator</u>. In my experience, the important distinction is often just whether a value is missing or present. I create new column[2] filled with a simple indicator that is 0 when the value is missing and 1 when the value is present.

   Note that it is critically important that you DO NOT simply "fix" missing values by hand, e.g., by using a text editor or the like. When you transition a model to a production scoring[3] environment, you

must be able to reproduce **everything** you did (typically automatically) when building the model – including handling missing values. This admonition applies to all the treatments presented in this section.

3. When a column has a significant number of missing values … Replace the missing value with a constant value. In addition to creating the missing/present indicator above, I also create a new column that replaces the missing value with a constant. For numeric columns, I typically replace the missing value with the mean or median. If the column is a non-numeric category, I replace the missing value with the mode (most frequent value). This kind of replacement normally has a fairly minimal impact on the distribution and statistics of the column.

4. When a column and its values are essential to producing accurate predictions/classifications … Estimate the missing value based on other, non-missing data elements. For instance, individual income may be a critical element of a loan-approval model, and it may occasionally be missing. You might estimate individual income using age, zip-level income estimates, and home value. Deciding that a column is critical is a judgment call, typically made AFTER a model has been developed. Also, since it can take significant effort to build a means to estimate the missing value, you will want to use this method very, very sparingly.

To summarize, I think it makes the most sense to work through the above options in the order listed. Find and use the **simplest** method(s) suited to the problem.

Intensive Care

There is an additional option to do something more in the scoring/production environment when a key value is missing. Consider a loan-approval model used to score new applicants. An applicant must receive a score above 0.8 to be approved. Notice that, for the applicant

AGE	EstIncome	LargeBal	TotalBal	Score
47		36	108	???

AGE	EstIncome	LargeBal	TotalBal	Score
47	22500	36	108	0.105

AGE	EstIncome	LargeBal	TotalBal	Score
47	150000	36	108	0.513

Table 2 - Two ways to score the same applicant

described by the record in the top portion of Table 2, the income value is missing.

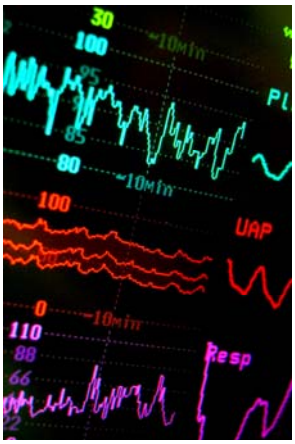
To make a decision regarding this applicant, we can do the following:

1. Fill the missing income value with the lowest value that occurred in the data used to build the model. Doing so produces a score of 0.105, as shown in the middle portion of Table 2.
2. Fill the missing income value with the highest value that occurred in the data used to build the model. This produces a score of 0.513, as shown in the bottom portion of Table 2.

Recall that the threshold for approval is 0.8. We see that even when we fill the missing value for this applicant with the best possible value, the score does not exceed 0.8, and we can conclude that the applicant should not be approved for a loan.

Post-Op

In this article on the nuts and bolts of data mining, we looked at missing values, one of the chief culprits behind "dirty data." I hope I motivated you to be concerned about missing values, since they can negatively impact analysis and predictive models. Beyond raising concern, however, I hope you also came away with a mental framework and the tools and techniques needed to detect, then diagnose, and finally prescribe treatments for managing missing values. Your data mining projects will be much more successful for your efforts!



Tim Graettinger, Ph.D., is the President of Discovery Corps, Inc. (http://www.discoverycorpsinc.com), a Pittsburgh-area company specializing in data mining, visualization, and predictive analytics.

Your comments and questions about this article are welcome. Please contact Tim at (724)-743-3642 or tgraettinger@discoverycorpsinc.com

[1] I've often seen periods and question marks, among others.

[2] I always like to keep the original column around for future reference. It also helps in the event I goof up the code I write to create the new column. Not that that's ever happened.

[3] Scoring is the term I use for the process of using the model to make predictions or classifications on new data, that is, data not used to develop the model. Note that the model is fixed and unchanging during the scoring process.