HARTE HANKS

A Refresher on Continuous Versus Discrete Input Variables

By: Sam Koslowsky, Senior Analytic Consultant, Harte Hanks

Continuously or 'Discreetly'

How many times have I heard that the most critical element in predictive analytics is the data? Don't misunderstand what I am saying. Method counts as well. But if there is a choice between better data or more methods, you can be sure that a data scientist would favor the richness of a data set.

Data comes in different formats. But we are able to classify data into two types- continuous and discrete. Bottom line is, if a variable can assume any value between its minimum and maximum value, then it is called a continuous variable. Much of the data we deal with fall in this category: age, income, spending, are all examples that we are most familiar with.

Pretty much all other variable types fall into the 'discrete' category. These can further be divided into categorical and ordinal.

A categorical variable is one that has two or more categories, but there is no intrinsic ordering to the categories. For example, a binary variable (such as gender-male/female) is a categorical variable having two categories and there is no intrinsic ordering to the categories. Car manufacturer is also a categorical variable having a number of categories (GM, Ford, Toyota, etc.) and here also, there is no universally approved way to order these from highest to lowest. Categories can be assigned. However, we cannot order them.

If a variable has a clear ordering, we refer to it as an ordinal variable. Examples of ordinal variables include: socio economic status ("low income","middle income","high income"), education level ("high school","BS","MS","PhD"), income level ("less than 50K", "50K-100K", "over 100K"), satisfaction rating ("extremely dislike", "dislike", "neutral", "like", "extremely like").

While there is a sense of ordering, one should observe the differences between adjacent categories do not necessarily have the same meaning. For example, the difference between the two income levels "less than 50K" and "50K-100K" does not have the same meaning as the difference between the two income levels "50K-100K" and "over 100K".

We often convert continuous variables into discrete ones. We do this by splitting up the continuous variable into ranges of values, or as they are frequently referred to as 'bins'. We then may assign the same discrete measure to all values of the continuous variable that fall within a certain range. For instance, income (a continuous variable) between $61,000 and $80,000 will be assigned a value of, say $70,000-a midpoint measure. Alternatively, it may just be categorized as falling in that range-between "$61,000-$80,000". In fact, it is not uncommon at all for us to "discretize" continuous variables and represent them in a discrete fashion.

Of course, many variables, have to remain in their original state. Gender, for example, is male or female. This piece of data did not originate from any continuous state. And typically, it is assigned to a 'dummy' variable format- '0' may represent male and '1' denotes male.

In developing analytic approaches, both for more intricate studies and model development, the analyst must decide whether to use a variable in its continuous state, or discretize it, and employ the variable in that form.

There is a common question among data scientists. Does it make sense to convert continuous variables to discrete ones? Take for instance the variable age. Is it better to leave it as a continuous variable, or to chop it into categories, e.g., 30 to 39, 40 to 49, 50 to 59, etc.? Will the continuous version of the data produce superior results, or will the binned data generate a better outcome?

I'm not certain there is an absolute answer to this question. It may depend on what your objective is. For example, are you an analyst involved in market research activities? Then you may very well employ categorical variables in your work This is the typical form that an audience may be comfortable with.

Another researcher, developing predictive models, may feel that expressing age as a continuous variable as a potential predictor, may provide additional insight.

It is fair to ask, "is there some analytic reason that might motivate an analyst to discretize, rather than use the continuous version of the variable?" Doesn't one lose information by chopping up the data? Suppose we have spending that refers to a customer's lifetime activity at a retailer. Spending is discretized into 'levels', so that we have HIGH, MEDIUM, and LOW categories. Some managers, subjectively assigning such ranges, may believe that something significant occurs at the cutoff. Does this make sense? Are these cutoffs correct? What happens if we modify the definitions, and associated ranges? Will our results change? They may, and this could lead the analyst to divide the continuous variable in a way that forces the results to conform to what one wants to see. It is poor practice to repetitively attempt to use different cut points of a continuous variable to secure a statistically significant result. We don't want to encourage the, "How to lie with Statistics" fallacy.

Suppose you are predicting profitability of some customer. If you bin income at 45-54k, 55-64k, 65=74k and 75k+, then you are implicitly assuming that a $58,000 income customer is more similar to a 55k customer than a 65k person. Something is wrong with the logic, here. Categorization assumes that the relationship between the predictor and the outcome is identical within intervals. This assumption, at the very least, is very questionable.

If age was discretized to “young” and “old” at say 46 years, then it is probable that pertinent information has been lost, essentially discarded. Two categories are probably inadequate, and we certainly do not want to remove any substantial learnings.

If the variable in question has a linear association with the result, some information is lost by discretizing a continuous variable. Additionally, if you constructed, say, six categories, you may have to estimate six coefficients, potentially generating a more complex model.

However, if the association is not linear, then the categories may allow you to capture the linear component of the relationship by pinpointing the category that appears to be significant. Treating the variable as continuous allows you to identify a potential linear relationship, but the discretized form may allow the analyst to locate more nuanced relationships-a beneficial feature.

Ok, that all makes sense. But how about examining ordinal variables and converting those to continuous? So, if we have five categories of income from low to high, labeled ‘1’ to ‘5’, we employ this piece of data as a continuous one. This tactic provides maximum flexibility in the approach of your analysis, and maintains the information in the ordering. Perhaps more critical to many data scientists, is that it allows one to analyze the data using techniques that the audience is more comfortable with, and which is more easily comprehended. The thinking being, that even if results are estimates, they’re probably reasonable assessments of what is going on. I will always examine an ordinal piece of data, and determine whether using it in a continuous form produces incremental value. It often does!

Another issue is one of interpretability. For example, suppose one computes the odds ratio for profitability for customers with income > $60,000 compared to persons with income below $60,000. The explanation of the resulting odds ratio is contingent on the distribution of incomes in the analysis sample (the proportion of subjects > $60k, <60k, etc.).

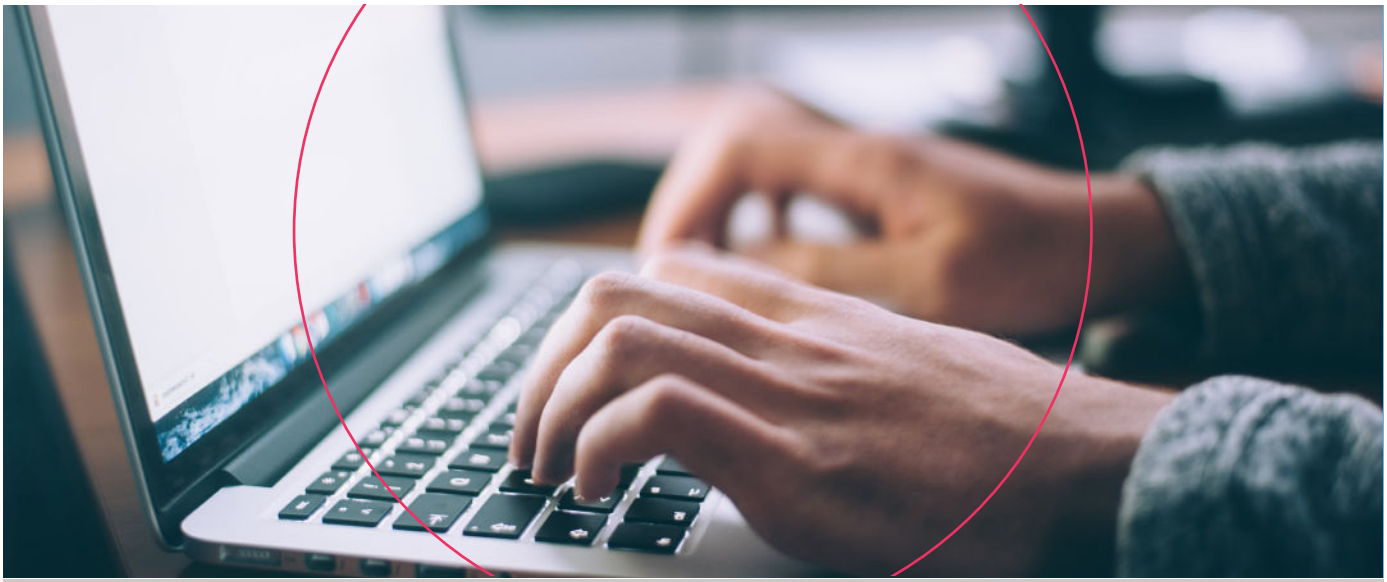
On the other hand, if profitability is modeled as a continuous variable one can estimate the ratio of odds for precise values of the predictor, e.g., the odds ratio for $135,000 income as compared to $60,000 income, thus providing increased understanding.

So, if I was going to discretize or 'bin' my continuous data, how would I go about doing it? While this is a legitimate discussion on its own, let me list a couple of methods that I have employed.

Equal record count results in a number of bin intervals based on the number of records being analyzed. Equal width binning is perhaps the most prevalent means of developing categories. After the binning, all bins have equal width, or represent an equal range of the original variable values, no matter how many cases are in each bin. Let's not forget, constructing bins based on the nodes that are produced through a decision TREE analysis. And finally, you can always find 'optimal' binning routines in a variety of analytic software.

So, is it good or bad to discretize?

Bottom line-if you must use discrete data, do it discreetly!



Contact us

www.hartehanks.com/contact