CoolData blog

22 September 2014

What predictor variables should you avoid? Depends on who you ask

Filed under: Best practices, Correlation, Model building, Pitfalls, predictive modeling, Predictor variables — Tags: data mining, predictive models, Predictor variables — kevinmacdonell @ 3:44 pm
People who build predictive models will tell you that there are certain variables you should avoid using as predictors. I am one of those people. However, we disagree on WHICH variables one should avoid, and increasingly this conflicting advice is confusing those trying to learn predictive modeling.

The differences involve two points in particular. Assuming charitable giving is the behaviour we're modelling for, those two things are:

1. Whether we should use past giving to predict future giving, and
2. Whether attributes such as marital status are really predictors of giving.

I will offer my opinions on both points. Note that they are opinions, not definitive answers.

1. Past giving as a predictor

I have always stressed that if you are trying to predict "giving" using a multiple linear regression model, you must avoid using "giving" as a predictor among your independent variables. That includes anything that is a proxy for "giving," such as attendance at a donor-thanking event. This is how I've been taught and that is what I've adhered to in practice.

Examples that violate this practice keep popping up, however. I have an email from Atsuko Umeki, IT Coordinator in the Development Office of the University of Victoria in Victoria, British Columbia*. She poses this question about a post I wrote in July 2013:

"In this post you said, 'In predictive models, giving and variables related to the activity of giving are usually excluded as variables (if 'giving' is what we are trying to predict). Using any aspect of the target variable as an input is bad practice in predictive modelling and is carefully avoided.' However, in many articles and classes I read and took I was advised or instructed to include past giving history such as RFA*, Average gift, Past 3 or 5 year total giving, last gift etc. Theoretically I understand what you say because past giving is related to the target variable (giving likelihood); therefore, it will be biased. But in practice most practitioners include past giving as variables and especially RFA seems to be a good variable to include."

(* RFA is a variation of the more familiar RFM score, based on giving history — Recency, Frequency, and Monetary value.)

So modellers-in-training are being told to go ahead and use 'giving' to predict 'giving', but that's not all: Certain analytics vendors also routinely include variables based on past giving as predictors of future giving. Not long ago I sat in on a webinar hosted by a consultant, which referenced the work of one well-known analytics vendor (no need to name the vendor here) in which it seemed that giving behaviour was present on both sides of the regression equation. Not surprisingly, this vendor "achieved" a fantastic R-squared value of 86%. (Fantastic as in "like a fantasy," perhaps?)

This is not as arcane or technical as it sounds. When you use giving to predict giving, you are essentially saying, "The people who will make big gifts in the future are the ones who have made big gifts in the past." This is actually true! The thing is, you don't need a predictive model to produce such a prospect list; all you need is a list of your top donors.

Now, this might be reassuring to whomever is paying a vendor big bucks to create the model. That person sees names they recognize, and they think, ah, good — we are not too far off the mark. And if you're trying to convince your boss of the value of predictive modelling, he or she might like to see the upper ranks filled with familiar names.

I don't find any of that "reassuring." I find it a waste of time and effort — a fancy and expensive way to produce a list of the usual suspects.

If you want to know who has given you a lot of money, you make a list of everyone in your database and sort it in descending order by total amount given. If you want to *predict* who in your database is most likely to give you a lot of money in the *future*, build a predictive model using predictors that are associated with having given large amounts of money. Here is the key point … if you include "predictors" that mean the same thing as "has given a lot of money," then the result of your model is not going to look like a list of future givers — it's going to look more like your historical list of past givers.

Does that mean you should ignore giving history? No! Ideally you'd like to identify the donors who have made four-figure gifts who really have the capacity and affinity to make six-figure gifts. You won't find them using past giving as a predictor, because your model will be blinded by the stars. The variables that represent giving history will cause all other affinity-related variables to pale in comparison. Many will be rejected from the model for being not significant or for adding nothing additional to the model's ability to explain the variance in the outcome variable.

To sum up, here are the two big problems with using past giving to predict future giving:

1. The resulting insights are sensible but not very interesting: People who gave before tend to give again. Or, stated another way: "Donors will be donors." Fundraisers don't need data scientists to tell them that.
2. Giving-related independent variables will be so highly correlated with giving-related dependent variables that they will eclipse more subtle affinity-related variables. Weaker predictors will end up getting kicked out of our regression analysis because they can't move the needle on R-squared, or because they don't register as significant. Yet, it's these weaker variables that we need in order to identify new prospects.

Let's try a thought experiment. What if I told you that I had a secret predictor that, once introduced into a regression analysis, could explain 100% of the variance in the dependent variable 'Lifetime Giving'? That's right — the highest value for R-squared possible, all with a single predictor. Would you pay me a lot of money for that? What is this magic variable that perfectly models the variance in 'Lifetime Giving'? Why, it is none other than 'Lifetime Giving' itself! Any variable is perfectly correlated with itself, so why look any farther?

This is an extreme example. In a real predictive model, a predictor based on giving history would be restricted to giving from the past, while the outcome variable would be calculated from a more recent period — the last year or whatever. There should be no overlap. R-squared would not be 100%, but it would be very high.

The R-squared statistic is useful for guiding you as you add variables to a regression analysis, or for comparing similar models in terms of fit with the data. It is not terribly useful for deciding whether any one model is good or bad. A model with an R-squared of 15% may be highly valuable, while one with R-squared of 75% may be garbage. If a vendor is trying to sell you on a model they built based on a high R-squared alone, they are misleading you.

The goal of predictive modeling for major gifts is not to maximize R-squared. It's to identify new prospects.

2. Using "attributes" as predictors

Another thing about that webinar bugged me. The same vendor advised us to "select variables with caution, avoiding 'descriptors' and focusing on potential predictors." Specifically, we were warned that a marital status of 'married' will emerge as correlated with giving. Don't be fooled! That's not a predictor, they said.

So let me get this straight. We carry out an analysis that reveals that married people are more likely to give large gifts, that donors with more than one degree are more likely to give large gifts, that donors who have email addresses and business phone numbers in the database are more likely to give large gifts … but we are supposed to ignore all that?

The problem might not be the use of "descriptors," the problem might be with the terminology. Maybe we need to stop using the word "predictor". One experienced practitioner, Alexander Oftelie, briefly touched on this nuance in a recent blog post. I quote, (emphasis added by me):

"Data that on its own may seem unimportant — the channel someone donates, declining to receive the mug or calendar, preferring email to direct mail, or making 'white mail' or unsolicited gifts beyond their sustaining-gift donation — can be very powerful when they are brought together to paint a picture of engagement and interaction. **Knowing who someone is isn't by itself predictive (at best it may be correlated)**. Knowing how constituents choose to engage or not engage with your organization are the most powerful ingredients we have, and its already in our own garden."

I don't intend to critique Alexander's post, which isn't even on this particular topic. (It's a good one — please read it.) But since he's written this, permit me scratch my head about it a bit.

In fact, I think I agree with him that there is a distinction between a behaviour and a descriptor/attribute. A behaviour, an action taken at a specific point in time (eg., attending an event), can be classified as a predictor. An attribute ("who someone is," eg., whether they are married or single) is better described as

a *correlate*. I would also be willing to bet that if we carefully compared behavioural variables to attribute variables, the behaviours would outperform, as Alexander says.

In practice, however, we don't need to make that distinction. If we are using regression to build our models, we are concerned solely and completely with correlation. To say "at best it may be correlated" suggests that predictive modellers have something better at their disposal that they should be using instead of correlation. What is it? I don't know, and Alexander doesn't say.

If in a given data set, we can demonstrate that being married is associated with likelihood to make a donation, then it only makes sense to use that variable in our model. Choosing to exclude it based on our assumption that it's an attribute and not a behaviour doesn't make business sense. We are looking for practical results, after all, not chasing some notion of purity. And let's not fool ourselves, or clients, that we are getting down to causation. We aren't.

Consider that at least some "attributes" can be stated in terms of a behaviour. People get married — that's a behaviour, although not related to our institution. People get married and also tell us about it (or allow it to be public knowledge so that we can record it) — that's also a behaviour, and potentially an interaction with us. And on the other side of the coin, behaviours or interactions can be stated as attributes — a person can be an event attendee, a donor, a taker of surveys.

If my analysis informs me that widowed female alumni over the age of 60 are extremely good candidates for a conversation about Planned Giving, then are you really going to tell me I'm wrong to act on that information, just because sex, age and being widowed are not "behaviours" that a person voluntarily carries out? Mmmm — sorry!

Call it quibbling over semantics if you like, but don't assume it's so easy to draw a circle around true predictors. There is only one way to surface predictors, which is to take a snapshot of all potentially relevant variables at a point in time, then gather data on the outcome you wish to predict (eg., giving) after that point in time, and then assess each variable in terms of the strength of association with that outcome. The tools we use to make that assessment are nothing other than correlation and significance. Again, if there are other tools in common usage, then I don't know about them.

Caveats and concessions

I don't maintain that this or that practice is "wrong" in all cases, nor do I insist on rules that apply universally. There's a lot of art in this science, after all.

Using giving history as a predictor:

- One may use some aspects of giving to predict outcomes that are not precisely the same as 'Giving', for example, likelihood to enter into a Planned Giving arrangement. The required degree of difference between predictors and outcome is a matter of judgement. I usually err on the side of scrupulously avoiding ANY leakage of the outcome side of the equation into the predictor side — but sure, rules can be bent.
- I've explored the use of very early giving (the existence and size of gifts made by donors before age 30) to predict significant giving late in life. (See Mine your donor data with this baseball-inspired analysis.) But even then, I don't use that as a variable in a model; it's more of a flag used to help select prospects, in addition to modeling.

Using descriptors/attributes as predictors:

- Some variables of this sort will appear to have subtly predictive effects in-model, effects that disappear when the model is deployed and new data starts coming in. That's regrettable, but it's something you can learn from — not a reason to toss all such variables into the trash, untested. The association between marital status and giving might be just a spurious correlation — or it might not be.
- Business knowledge mixed with common sense will help keep you out of trouble. A bit of reflection should lead you to consider using 'Married' or 'Number of Degrees', while ignoring 'Birth Month' or 'Eye Colour'. (Or astrological sign!)

There are many approaches one can take with predictive modeling, and naturally one may feel that one's chosen method is "best". The only sure way to proceed is to take the time to define exactly what you want to predict, try more than one approach, and then evaluate the performance of the scores when you have actual results available — which could be a year after deployment. We can listen to what experts are telling us, but it's more important to listen to what the data is telling us.

//////////

Note: When I originally posted this, I referred to Atsuko Umeki as "he". I apologize for this careless error and for whatever erroneous assumption that must have prompted it.

Comments Off