Predictive Analytics and Futurism
December 2016
Issue 14

Creating a Useful Training Data Set for Predictive Modeling

By Anders Larson

In the past decade, the capabilities of predictive analytics have improved dramatically thanks to greater availability of large data sources, increased computing power, and innovation from the statisticians, data scientists, and actuaries at the forefront of the field. As a result, there has been more and more interest from companies across nearly every industry to harness the power of machine learning and other advanced predictive modeling techniques.

For all the advancements that have been made, the ability to produce useful and accurate results with any of these techniques is still ultimately reliant on one thing: robust and appropriate data with which to train the model. This goes beyond the simple "garbage in, garbage out" principle. There's no doubt that data with blatantly incorrect or sparsely populated information won't do us much good in building a predictive model. It should go without saying that data cleaning is an essential step in the model building process.

We could train a model with an immaculately clean data set with 500 million records and 100 variables, then use that model to make predictions using an equally clean data set with the exact same set of 100 variables, and we could still end up with awful predictions if the model is based on faulty assumptions. In fact, this is perhaps one of the most dangerous situations, when it seems for all the world like we have a model we can trust, and so we do trust it, until it's too late, when it becomes clear that our predictions were just … bad.

One of the most important elements of a useful training data set is that it is a reasonable representation of the data we'll be using to make predictions about the future. In general, we want the same data generating process underlying the training data to plausibly apply to any new data fed to the model when making predictions. Even with clean data, there are often subtle biases in training data that can cause us to build a model that is inappropriate to apply to new data. To help provide more clarity, I'm going to describe a few specific examples in more detail. What I hope to accomplish here is to heighten awareness, so that, the next time you begin building a training data set, you can be on the lookout for the dangers that might be hiding in the data. I apologize in advance for my own not-so-subtle bias toward examples from the health care world.

LIMITATIONS ON PATIENTS WITH CLAIMS HISTORIES

Accountable care organizations (ACOs) participating in the Medicare Shared Savings Program (MSSP) receive full claims detail from the Centers for Medicare and Medicaid Services (CMS) on all of the patients assigned to it. In an ACO's first year, CMS provides claims histories for all currently assigned patients, extending back one year prior to the start of the ACO. The data is clean, consistent, and reliable, and given that all ACOs should have at least 5,000 patients, there are plenty of observations with which to train a predictive model. However, there is one catch: CMS does not provide any claims history for patients who died prior to the start of the ACO's first year (decedents).

For most Medicare ACOs, approximately 5 percent of patients alive at the start of the year will die by the end of the year, and

DEFINING KEY TERMS

Training data set—The set of records used to calibrate a predictive model and determine relationships between characteristics (features) and a particular outcome (response). The data should be divided into two subsets, often based on a time or date variable. The two subsets are:

- **Feature:** The data used to gather characteristics that will be used as features in the model. For instance, in a model that predicts health care costs, this time period would be used to determine things like clinical conditions, historical costs, and historical utilization of services. In a standard, prospective predictive model, this data would come from an earlier time period than the response set described below.
- **Response:** The data used to observe the outcome you are looking to predict. For purposes of calibrating the model, this data set should not be used to determine any of the characteristics that will be used as features. In a standard, prospective predictive model, this data would come from a later time period than the feature set described above.

Prediction data set: The set of records used to make new predictions using the model calibrated on the training data set. This data set is similar to the training feature data, but likely for a more recent time period (often the most recent time period). It is used to gather the same characteristics as the feature portion of the training data above. ■

these patients generally incur very high costs in the last few months of life.[1] When constructing a training data set, it would make sense to remove patients who died during the feature period, because there would be no need to predict their future costs. However, you would want to include patients who died during the response period, because it is likely that some patients for whom you will make predictions will die in the predictions period.

Unfortunately, in this situation we have no decedents available in the training response period, which creates a bias in the training data set. The patients selected for inclusion in the training data set are, on average, healthier and lower-cost than the patients for whom you will be making predictions. This is true even after accounting for other patient characteristics, such as the presence of chronic conditions. The average patient with congestive heart failure who does not die in the next six months is still much less costly than the average patient with congestive heart failure who does die in the next six months.

As a result, the predicted costs will be understated for a model trained on this data set. Communicating the limitations of the model to the end user will be particularly important in this situation (and complicated). The predictive models can still be quite useful, as long as the focus is more on the rank order of the predicted costs rather than the specific level of predicted costs. Conversely, using this model to predict population-level costs would likely be inappropriate. Ultimately, as time passes, the ACO will receive enough data on patients who die after the start of the performance year, and a more appropriate model can be retrained.

DIFFICULTIES WITH NEW AND EXPANDING POPULATIONS

The ever-evolving health care landscape in the United States presents opportunities for predictive modelers, but not without additional challenges. One situation that can be particularly tricky is when a new class of patient is introduced into a population. The expansion of the Medicaid program in many states is a particularly instructive example.

In most states, the Medicaid population prior to 2014 was comprised of disabled adults and low-income families and children. Under the Patient Protection and Affordable Care Act (ACA), states are encouraged to extend eligibility to low-income adults who did not otherwise qualify. The morbidity levels of these newly eligible patients was a huge unknown prior to the start of the program because many of these patients had been previously uninsured.

But let's take a step forward and look at the situation even after a year has passed since the expansion of Medicaid in a particular state. Assume you are constructing a predictive model to predict individual and population-level costs for a managed Medicaid plan. You can use the past year of history to build a training data set, but that training data set may have biases built into it. All of the patients who were newly eligible for Medicaid did not enroll immediately, and those who do enroll right away may not be representative of the type of patients who will ultimately enroll (for instance, the early enrollees may have higher morbidity on average or have pent-up demand for services).

Compounding this problem is the fact that there was a relatively small number of these patients in the training data set, and yet

> The ability to produce useful and accurate results with any of these techniques is still ultimately reliant on one thing: robust and appropriate data with which to train the model. This goes beyond the simple "garbage in, garbage out" principle.

the population now represents a much larger portion of the prediction data set. The predictive model you construct is going to extrapolate the learnings from that small sample to make predictions for a much larger group, which will exacerbate any biases you have in your training data set. As actuaries, exercising judgment in this situation is essential. Placing too much trust in a model that is not necessarily aware of outside influences, such as pent-up demand, is a serious risk. One option in this case would be to look at outcomes in other states that expanded Medicaid previously. When viewed at a population level, do the results of your predictive model look reasonable compared with experience elsewhere?

CHALLENGES WITH TRANSACTIONAL DATA

The challenge of creating a training data set becomes more complicated when dealing with transactional data. In the examples above, we have a way to measure the number of patients (or more generally, the exposure units) that are associated with the claims or other utilization measures that occur. In some cases, data is simply provided about transactions (claims, services, payments, etc.) as they occur, with no corresponding information about which patients were eligible to have these transactions.

Using enrollment or exposure information, rather than claims information, to select patients for the training data set generally makes the most sense. This enables an understanding of the difference between patients who could have used services and patients who did not use any services because they were not eligible (or were not included in the original data). With transactional data, such as an electronic health record, the predictive modeler often has to make an attempt to infer some type of exposure metric. One option is to look for the first date that a patient appeared in the electronic health record and assume that the patient was "eligible" to receive services from that point forward.

These inferred enrollment estimates will also be needed to select patients for inclusion in the training data set. In these situations, particular caution must be used to avoid biasing the training data set. In general, it is dangerous to use anything learned in the training response period to determine which records to include in the training data set. Rather, pretend that your response period is truly unknown, just like the future you're trying to predict. In this example, our best approach would be to include patients whose inferred enrollment began prior to the start of the training response period and were therefore "eligible" to receive services in the training response period.

This approach will still yield less-than-perfect results. For instance, assume the data provided to you includes all services in the past 24 months. Then assume there is a patient who had only one service in the data, and it occurred three months ago, but this patient also had a service 30 months ago, which is not in the data. Let's say you are training the model to predict services over a 12-month span, so you set the training response period to begin 12 months ago. This patient would be excluded from the training data set because it was assumed that person was not eligible to receive a service. Had the data been cut six months earlier, you would have observed that initial visit 30 month ago, and the patient would have been included in the training data set with no services in the training response period.

Unfortunately, there is no magic bullet for handling transactional data with no exposure information. Getting a thorough understanding of the data-generating process underlying the data will help, but it is critical to be aware of the limitations and potential uncertainty of a model built on this type of data.

CONCLUSION

The examples in this article are by no means exhaustive. Every predictive modeling scenario has its own unique challenges, and arguably it's never possible to put together a training data set that is a perfect representation of the prediction data set. But taking care to create a useful and appropriate training data set is an often underappreciated step in the predictive modeling process. There's no question that expertise in selecting and calibrating the model itself is a vital skill, as is the ability to communicate and interpret the results, but any model will be imperiled from the start without a solid understanding of the data used to train it. ■

Anders Larson, FSA, MAAA is an actuary at Milliman in Indianapolis. He can be reached at *Anders.larson@milliman.com.*

ENDNOTES

[1] Riley, G.F. & Lubitz, J.D. (April 2010). Long-Term Trends in Medicare Payments in the Last Year of Life. Health Services Research. Retrieved September 2, 2016, from *http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2838161/.*