Article from

Predictive Analytics and Futurism

July 2016
Issue 13

Bridging the Gap

By Bryon Robidoux

On Nov. 15, 2015, I attended Bridging the Gap Series: Application of Predictive Modeling in VA/FIA Risk Management at the Equity Based Guarantees Conference in Chicago. There were four major sections to this session: introduction/setting the stage, basics of generalized linear models (GLM), the case study and practical issues outside of building the predictive model. This article will be a review of the subjects covered in this session of the conference.

The introduction/setting the stage was probably the most disappointing part of the class. It only lasted for a half hour, but I thought most of the information had little to do with predictive modeling. It had more to do with different risk profiles of varying annuity products and how they relate to each other. Most people at this conference would be in the business and have a good handle on this information. The part related to predictive modeling was more common sense than informative. It could have been cut and nothing would have been missed.

The section on the basics of GLM was great. This section covered ordinary regression, gamma regression, link function, bias versus variance and dangers of collinearity. If a person had any aptitude for mathematics at all, he or she would be able to follow the demonstration. The slides were at the appropriate level to introduce everyone to the purpose of GLMs and give a sense of when and where you might use them. No details were stated that were not absolutely necessary. As actuaries, we constantly have to present technical information and we struggle with providing the appropriate level of detail to an audience. This presentation was a great example of how to exactly do that. I really enjoyed this part.

The case study section explored how to build a model for whether or not a policyholder would make a renewal deposit. The topics covered in the case study were

- log likelihood,
- data,
- applying GLMs,
- model selection,
- back testing,
- visualization of results,
- weighted data,
- adding interactions,
- non-categorical factors,
- individualized behavior,
- logistic regression,
- producing the final model.

It was a lot of information to cover in less than two hours. All the information was great and relevant, but it felt very rushed and I was overwhelmed very quickly. This may be why I retained very little of the lecture. It would have been more digestible if half

the topics were covered or if it had been an entire day. I will stop short of saying I wish this was a hands-on tutorial; however, I would have enjoyed the presenter showing the R or Python code written to create the tables used in the presentation. If the data and model had been published on Github.com, I could walk myself through the demonstration when I got back home. I would like to see this as a standard for demonstrations like this going forward. This may not be possible because the data may be proprietary, but I am hoping presenters will cleanse the data so this is not an issue.

I was disappointed with the material in the backing testing section. They really gave the audience the impression that splitting the data between training and test was to arbitrarily split the data 70/30, respectively. The approach the modeler uses to divide the data between training and test is a very important part of the modeling process, especially when the data is sparse. Data is a valuable resource and should be managed as such. There should have been a focus on cross validation techniques so the audience had a better understanding of how to split their data properly. The only other detail to nitpick is that the presenter was using confidence interval and prediction interval interchangeably. These are not the same and it is important to understand the difference.

First, a confidence interval and prediction interval are used in different contexts. A confidence interval is used when estimating a population model parameter θ. A prediction interval is used when predicting the outcome of a response random variable Y in a model. For example, prediction problems occur when you are interested in a gain from an investment made next month, rather than the mean gain over a long series of investments.[1] Mathematically the prediction and confidence intervals are very closely related and I think this is where the confusion arises. Assume we have a large amount of data, the (1-α)100% confidence interval is

$$(\vec{\theta} - z_{\frac{\alpha}{2}}\sigma_{\hat{\theta}} < \theta < \vec{\theta} + z_{\frac{\alpha}{2}}\sigma_{\hat{\theta}})$$

where θ is a point estimator for the parameter, σθ̂ is the standard deviation of the point estimator and Z is the distance from the mean measured in standard deviations from a normal distribution.

In a prediction, we are concerned with error in the actual versus predicted response. The (1-α)100% prediction interval is

$$P(\widehat{Y^*} - z_{\frac{\alpha}{2}}\sigma_{error} < Y^* < \widehat{Y^*} + z_{\frac{\alpha}{2}}\sigma_{error})$$

where Y* is the value of the actual response Y when the independent variable x is equal to a particular value x*, $\widehat{Y^*}$ is the predictor of Y*, and σ_{error} is the standard deviation in the error between the actual response Y* versus the predicted response $\widehat{Y^*}$. The variance of the error V(error) equals the variance of the actual

$$V(Y^*) + \text{the variance of the predictor } V(\widehat{Y^*}).$$

The key concept is that the predictor $\widehat{Y^*}$ can be viewed as just another point estimator θ̂. Mathematically the only difference between the prediction interval and the confidence interval is in the variance, such that the variance of the prediction interval needs to include the variance in the actual response. It is this additional amount of variance above the variance of the point estimator that always makes the prediction interval wider than the confidence interval.

The last section of the day was about practical issues outside of building a predictive model. The focus of this section was on communication. The presenters had some very good points and it is worth restating them.

As the decision moves down the management ladder, the decision-maker will ask some fundamental questions:

1. What can predictive modeling do for us?
2. Where should we apply predictive modeling ?
3. What data should be provided to the predictive model?
4. What should our predictive model be?

Question 1 is concerned with getting senior management to see the importance of predictive modeling and being able to provide them with benchmarks to show how predictive modeling helps the bottom line. With all the hype of predictive modeling, it is also concerned with managing senior managements' expectations on what can be reasonably accomplished. Right now, they may think it is the panacea for all that ails the business.

Question 2 is concerned with when it is appropriate to build a model and whether or not the cost of building the model is worth the insight that will be achieved. They stated the hazard of predictive modeling increases with

- modeling severity and not just frequency,
- high correlation among potential and explanatory factors, and
- most importantly, the lack of sufficient and directly applicable data.

Question 3 is concerned with the difficulty of retrieving the data for the model. Is the data internal or external? How often does the data remain relevant? Is the data grouped? Are manual processes required to assemble the data?

Another theme in the presentation was the role of the actuary in predictive modeling. The presenter shared an analogy, which I will paraphrase: "Just because anyone in the audience can go on-

line and learn how to give a root canal, doesn't mean I am going to allow anyone in the audience to give me one." His statement resonated with me on multiple levels:

1. What does it mean to become something, such as an actuary, data scientist or software developer?
2. What is the proper communication between the data scientist and the actuary?
3. What are the responsibilities of the data scientist versus the actuary?

I have been watching "Comedians in Cars Getting Coffee," a funny webcast by Jerry Seinfeld. One of the major objectives of the show is to break down what it means to be a comedian. I find it interesting that they always think a person is born a comedian and it can't be learned, but they proceed to share how they stunk in the beginning and hard work and multiple shows daily got them where they are today.

As I try to get into predictive modeling, I have been struggling with what it means to be a data scientist. To be honest, some days I struggle with what it means to be an actuary. What I have determined is that every profession has an art and a science. The science can be learned by reading books and taking exams. The art can only be learned in the trenches by spending a large majority of each day focused specifically on solving problems in the professional domain. While taking an exam or a class to learn the science, the goal is to get the correct answer to the presented problem. To master the art of a profession, the goal is to learn how to fail. Both newbies and professionals will fail, but the professional will know how to analyze the failure and turn it into success.

For this reason, I agree with the presenters that, in most cases, it doesn't make sense for the actuary to become a data scientist. Predictive modeling is a huge topic and there is a ton of art to being a data scientist or statistician! It is easy to learn linear regression and to get a basic understanding of GLMs, but this is a long way from building a truly usable model. It is one thing to go through the examples in a book. It is another thing to have a supervisor plop a couple of files in a directory with sparse documentation and tell you to build a model in one week for a presentation for her supervisors.

One presenter said the role of the actuary in predictive modeling is to instill the business knowledge into the data scientist. It is not for the actuary to become the data scientist. A data scientist will look at the data and try to find the best model. They might find inputs are strongly correlated with the response, but the model may not make complete sense from an actuarial or business perspective. It is the job of the actuary to explain the business to the data scientist so he or she can more effectively do their job. The better the communication between the two parties, the better the end result will be.

At RGA, we have a brilliant mathematician/data scientist in my area. We wanted him to build us a model to better understand our lapses and withdrawal utilization. We were a little disappointed that the work product was just a little more than the actual versus expected analysis. We felt we could have easily produced the information ourselves. We were frustrated that we were not getting more informative insight from him. This presentation made me realize the problem was not with the mathematician but with me! It is very easy to point fingers. All we did was plop our raw data on his desk and ask him to build us some models. I did not enlighten him on the background information he needed. With a little work, I could have transformed the data and injected additional data so the fields were more representative of the problems to be solved. I could have taught him the relative information he needed to be more successful. It is a poor excuse to say that I was too busy on other projects and didn't have time to help. Now that I have accepted responsibility, we are getting much better results.

In conclusion, I thought Bridging the Gap Series: Application of Predictive Modeling in VA/FIA Risk Management was worthwhile to attend. I thought all the information provided was relevant to predictive modeling. There is no reason that only variable annuity (VA) or fixed indexed annuities (FIA) actuaries should have attended. It was applicable to a wider audience. Actually, I wish it would been a little more tailored to FIA and VA concerns, such as utilization and dynamic lapse. I also wish the case study portion was slowed down and lengthened. It would have helped solidify the information. Lastly, I liked that the presentation ended talking about communication. It is important to consider the best way for actuaries to communicate with statisticians/data scientists and how actuaries should communicate with their management about predictive models. ■

Bryon Robidoux, FSA, is director and actuary, at AIG in Chesterfield, Mo. He can be reached at *Bryon.Robidoux@aig.com.*

ENDNOTES

1 Dennis D. Wackerly, William Mendehall III and Richard L. Scheaffer, "Predicting a Particular Value of Y Using Simple Linear Regression," ch. 11.7, in *Mathematical Statistics with Application*, 5th ed. (Belmont: Duxbury, 1996), 506-09.

2 Ibid.