SOCIETY OF ACTUARIES

PREDICTIVE ANALYTICS AND FUTURISM SECTION

Predictive Analytics and Futurism

ISSUE 13 • JULY 2016

Predictive Modeling Techniques Applied to Quantifying Mortality Risk

By Vincent J. Granieri

Predictive Analytics and Futurism

Issue Number 13 • JULY 2016

Published by The Predictive Analytics and Futurism Section Council of the Society of Actuaries

475 N. Martingale Road, Suite 600
Schaumburg, Ill 60173-2226
Phone: 847.706.3500 Fax: 847.706.3599

SOA.ORG

This newsletter is free to section members. Current issues are available on the SOA website (www.soa.org).

To join the section, SOA members and non-members can locate a membership form on the Predictive Analytics and Futurism Section Web page at http://www.soa.org/predictive-analytics-and-futurism/.

Change or Be Changed!

By Dave Snell

This is my eighth year editing our section newsletter. Frankly, I am as excited by this opportunity to spread the knowledge of our many talented members today as I was in 2009 when we resurrected the newsletter of the former Futurism Section. It had not been published in quite a while. Back then, it was a struggle to get enough articles for an issue. Four of us wrote seven articles so we could start publishing again.

This issue has 17 articles contributed by 18 authors. They cover many topics we never saw in actuarial study notes. In fact, some of our authors, with SOA board endorsement, are helping to get this new material onto the syllabus for future actuaries. This level of sharing, and advancing the profession, is an excellent way to keep the actuary viable in a rapidly changing world—a world that has seen many professions diminish in stature and economic practicality. We see taxi service replaced by Uber and Lyft, and soon by self-driving cars. An Oxford University study[1] of 702 occupations showed insurance underwriters in 698th place, with a 99 percent probability of computerization. They were safer from obsolescence by automation than only four occupations, including hand sewers and telemarketers.

Several authors talk about how important it is to change—as if that were easy. Yes, we humans often like to initiate change, but we tend to be somewhat change averse when we are the ones affected. The only person who welcomes being changed is a baby with a wet diaper.

In the spirit of embracing the more recent tools and techniques available, we conducted a Delphi study to choose our name, the Predictive Analytics and Futurism Section (PAF). The Delphi study eliminates the biasing influence of hierarchy and involves rounds of anonymized responses that form the input for the successive rounds until study participants stop changing their minds. The name that emerged better reflected our focus, and it attracted a lot of new members. The increase in PAF membership over the past year is 65 percent, and we are happy to welcome so many new members.

The article topics continue to impress me, and it is a challenge to learn enough about a new topic to edit an article on it and still meet our deadlines. In that regard, I wish to introduce Kevin Jones as my associate editor for this issue (likely becoming co-editor for our next issue). Kevin is a brand-new FSA with a master's degree in mathematics and lots of modeling experience. He also is a winner (twice) of the Reader's Choice Award in the Actuarial Speculative Fiction Contest, which we co-sponsor with the Technology and Actuary of the Future Sections. Please join me in welcoming Kevin to our expanded editorial staff. We hope to continue to bring you high quality articles for both beginners and experienced PAF practitioners.

Now, let's discuss the contents of this issue.

- **Chairperson's Corner**, by Brian Holland. Brian describes how our section membership has dramatically increased since the name change, and describes some of the major initiatives PAF has introduced or improved upon, including webcasts, seminars, SOA meeting sessions, podcasts, LinkedIn discussions and our newsletter. Read it and be proud of our many accomplishments! Also, check to see what areas Brian describes that you might have overlooked lately, or areas in which you can contribute more.

In the spirit of embracing the more recent tools and techniques available, we conducted a Delphi study to choose our name. ...

- **Predictive Modeling Techniques Applied to Quantifying Mortality Risk**, by Vincent J. Granieri. Vince describes the Cox proportional hazards model and how actuaries and underwriters use this to establish debit and credit values in the underwriting process. An advantage of this model, as Vince tells us, is that it can accommodate data where some subjects leave or die along the way, others enter part-way through, and others have multiple underwriting events. Read his insights from a study of more than 80,000 lives tracked for up to 15 years through 200,000 underwriting events.

- **2036: An Actuarial Odyssey with AI**, by Dodzi Attimu and Bryon Robidoux. In this article, Dodzi and Bryon coin a new term: AI-calypse—the merger of artificial intelligence and an apocalypse. Will the continued progress of AI and machine learning lead to more prosperity for humanity or will the impact be negative? Read as they describe the Robo Actuary and the Robo Actuarial Analyst, what these new players might do in a typical day, and the impact this may have on the actuarial

CONTINUED ON PAGE 5

Chairperson's Corner

By Brian Holland

Did you know our section membership increased by about 65 percent in the last year? The council debated whether to change the section name (and what would be more indicative of our section), and it seems that now more interested parties are finding us. Some actuaries are likely still renewing their SOA membership and the tally might grow further still.

It just goes to show you that these times are exciting indeed for actuaries with an interest in predictive analytics. This issue I would like to welcome our many new members and review section activities.

THIS NEWSLETTER

We consider the newsletter our crown jewel. It appears semiannually. Please explore past issues (*https://www.soa.org/news-and-publications/newsletters/predictive-analytics-and-futurism/default.aspx)*. As you do, you will notice we were interested in predictive analytics long before the change of the section's name.

WEBCASTS

Our webcasts are often cosponsored with other sections to gain a wider audience, as only sponsoring section members are informed about the webcasts.

PRACTICAL PREDICTIVE ANALYTICS SEMINAR

A major new step for us is a one-day seminar after the Life & Annuity Symposium in Nashville in May. By print time, it will be done. We have aimed for a larger group—up to 50—and our goal is to apply predictive analytics in the life and annuity space. Learning techniques online is fine, but actuaries as a community must adopt practices and see the meat of what they are in an actuarial context, hence this seminar. In line with my last chairperson's corner, we reached out and brought in a data scientist to speak for part of the day.

SOA SESSIONS

This year we are sponsoring even more SOA sessions, including two at the Valuation Actuary Symposium, to be held in Hollywood, Fla., in August. ValAct seems to me to be a major application for predictive analytics, or setting assumptions, in other words. I expect there will be even more sessions at future ValActs.

PODCASTS

As of this writing, we have two podcasts on the SOA website (*https://www.soa.org/Professional-Development/Event-Calendar/Podcasts/Predictive-Analytics-and-Futurism.aspx)*: one on machine learning and one on the bias-variance trade off. I won't summarize them. Just listen to them. They're about 20 minutes each. Thanks go to your section council members Shea Parkes and Anders Larson for their recordings.

LINKEDIN GROUP

On LinkedIn, we have a forum (*https://www.linkedin.com/groups/5118314*) for online discussion, sharing links and the like. Few people have signed up as yet. Fix that, please.

Fortunately for us, several friends of the council continue to perform substantial work to get content out in front of you. Dave Snell remains on as our newsletter editor and is now assisted by Kevin Jones, another friend of the council. Also, Richard Xu and Dorothy Andrews are leading our sessions at the Annual Meeting & Exhibit in Las Vegas in October and ValAct respectively. ■

Brian D. Holland, FSA, MAAA, is director and actuary, Individual Life and A&H Experience Studies at AIG. He also serves as chair of the Predictive Analytics and Futurism Section Council. He can be reached at *brian.holland@aig.com.*

profession. Fans of *2001: A Space Odyssey* will see the dangers they describe. They urge us to reinvent ourselves and avoid being minimized and potentially eliminated.

- **Beyond Multiple Regression**, by Michael Niemerg. Michael's summary paragraph states, "There are many sophisticated models and methods beyond multiple regression that can be useful to a modeler," and his article describes some enhancements such as the least absolute shrinkage and selection operator (LASSO), ridge and least angle regression (LARS) models. He helps clear some of the confusion about when and where to use them to avoid classic multiple regression issues such as overfitting the data and overestimating the impact of variables of small effect. Yes, formulas are included, but they are explained and he provides visualization charts to help the learning process.

- **An Introduction to Incremental Learning**, by Qiang Wu and Dave Snell. Qiang and I describe how machine learning is a natural enhancement to predictive analytics through its many useful tools to derive meaning from a lot of otherwise unusable data such as handwriting. Again, there are formulas involved but we have tried to keep them understandable to the actuary who isn't already familiar with stochastic gradient descent, perceptrons and principal component analysis. Incremental learning can help you refine the estimations you derive from more traditional actuarial methods.

- **Follow Your Passion**, by Shea Parkes. Shea describes his journey from "bumbling beginner" (difficult to imagine for those of us who know Shea) to "intraprenaur." He gives hope to those who want to become more entrepreneurial without leaving their infrastructure and moving to Silicon Valley. You don't have to live on wheat grass and technobabble while trying to find your niche. Shea provides tips on how he combined his passion for learning with his actuarial training and experience to create a career path for himself.

- **Bridging the Gap**, by Bryon Robidoux. As you undoubtedly have read from past issues, we encourage actuaries to continue their learning process and stress that much of that learning should be at conferences beyond the usual actuarial meetings. Bryon summarizes his experience attending Bridging the Gap Series: Application of Predictive Modeling in VA/FIA Risk Management, and he gives us some insightful (and amusing) thoughts about what it means to be a data scientist and, for that matter, what it means to be an actuary! His description of the art versus the science is worth perusal and thought.

- **An Insurance Company for the 21st Century: A Thought Experiment**, by Jeff Huddleston and Benjamin Smith. Ever wonder how predictive analytics, if implemented throughout an insurance company, could transform it into a lean, mean,

profitable machine? Combining electronic health records and the ubiquity of data about an applicant from other sources, this business model might be possible, and their article is an intriguing look at a potential implementation.

- **Three Pitfalls to Avoid in Predictive Modeling**, by Marc Vincelli. Marc shows us we have a wonderful tool in predictive analytics; however, if we misuse that tool, there may be more damage than benefit. His three tips for effective use of predictive modeling techniques are sage advice for any actuary—especially those building or using predictive models.

- **The Actuarial Road Not Taken: Jeff Sagarin's Sports Ratings**, by Anders Larson. Some people work at a job to fund a future dream. Others, such as Jeff Sagarin, find a way to create a job that embodies the dream. All actuaries, but sports fans in particular, will enjoy reading Anders' account of how Jeff used his training as an actuary and his love of sports to merge the two into a job supplying USA Today and many other media outlets with better ratings than the opening betting lines.

- **Seasonality of Mortality**, by Kyle Nobbe. Kyle gives us a reality check. Yes, there are many extensions to simple linear regression, but actuaries have used this very powerful and time-tested technique to significant advantage. Kyle details a study of the seasonality of influenza and pneumonia, and their impact on other causes of death, such as diabetes and heart disease, using simple, but still elegant, regression.

- **Using Hadoop and Spark for Distributed Predictive Modeling**, by Dihui Lai and Richard Xu. The newer predictive modeling tools are powerful but they work best when supported by a lot of computing power. Parallel processing

is very powerful but often very difficult to implement. Dihui and Richard take us through the maze of options with H2O, SparkNet/MLib and Mahout. Which do you use for a generalized linear model (GLM), and which is better for high-volume data manipulation? Read their advice and save a lot of time, cost and effort.

- **Regression and Classification: A Deeper Look**, by Jeff Heaton. Supervised training almost always involves some forms of classification or regression (or both). The choice is usually based on whether the output will be discrete classes or a computed numerical value. Jeff explains these two pillars of predictive analytics and gives the reader a better grounding for model choices between GLMs, linear regression, neural networks, support vector machines and tree-based models. Learn how the receiver operating characteristic (ROC) curve can help you avoid overfitting, and lots more.
- **From Deep Blue to Deep Mind: What AlphaGo Tells Us**, by Haofeng Yu. The news media have made a big deal about the recent defeat of the world's reigning human Go champion by a computer program. Haofeng, a long-time player of Go, shares an insider's view of how impressive this feat is, and what implications it has for the future of machine learning capabilities. He also puts the win in perspective: None of the programs so far have exhibited general intelligence—the kind we associate with humans. Read his view of how these programs work and how they still have shortcomings.
- **Exploring the SOA Table Database**, by Brian Holland. A common complaint of data scientists is the need for more data. Brian describes a great source of mortality data from our own profession—the Society of Actuaries! The number of table files exceeded 2,600 but the number of dimensions involved for some of them make inferences across tables more complicated. Most of us have trouble imagining how to summarize a table with, say, 140 dimensions. Brian teaches us about dimension reduction techniques, such as singular value decomposition (SVD), and uses them to detect a valuation system error among the thousands of values.
- **The Impact of Deep Learning on Investments: Exploring the Implications One at a Time**, by Syed Danish Ali. Deep learning works well with unstructured data—the kind of data that other methods falter on—and it is rapidly being embraced by IBM, Google, Baidu and other companies seeking to capitalize on its potential for learning in a layered approach, as we humans do in facial recognition and handwriting analysis. Danish projects that it may also be suitable for something closer to an actuarial focus—the area of mergers and acquisitions (M&A).

As you can see, this issue is another great collection of articles—except we do not see the one from you! Please send us your ideas. If you are a little unsure about how to get started, Kevin and I are happy to help you through the process. Contribute! Grow! Be the change agent, rather than just the person being changed! ■

Dave Snell, ASA, MAAA, is technology evangelist at RGA Reinsurance Co. in Chesterfield, Mo. He can be reached at *dave@ActuariesAndTechnology.com.*

Kevin Jones, FSA, CERA, is associate actuary at Milliman in Buffalo Grove, Ill. He can be reached at *Kevin.Jones@Milliman.com.*

ENDNOTES

[1] Carl Benedikt Frey and Michael Osbourne, "The Future of Employment: How Susceptible are Jobs to Computerisation?" (paper, Oxford Martin School, University of Oxford, September 2013), *http://www.oxfordmartin.ox.ac.uk/downloads/academic/The_Future_of_Employment.pdf.*

Predictive Modeling Techniques Applied to Quantifying Mortality Risk

By Vincent J. Granieri

Actuaries are familiar with the interaction of art and science in their work. Some view underwriting in the same way, perhaps concluding that underwriting leans more toward art than science. With the advent of powerful computers and predictive modeling tools, it is possible to analyze survival data and produce statistically credible underwriting models that predict relative mortality risk among individuals based on demographic information and relevant conditions. In this article, we will discuss the use of the Cox proportional hazards model in developing a predictive underwriting model that produces a mortality multiplier for each individual. This multiplier can serve as the basis for debits and/or credits as it expresses the relative risk of having a given condition vis-à-vis not having it.

Further, we will attempt to quantify the impact on survival, if any, of being a member of certain subpopulations. We were looking to validate the time-accepted concepts of the wealth effect (in the wealthier subpopulations, which is beyond the scope of this paper) and antiselection (among insureds who sell their policies) in our population.

COX PROPORTIONAL HAZARDS MODEL

The Cox proportional hazards model was introduced in 1972 as a method to examine the relationship between survival (mortality) and one or more independent variables, called explanatory variables. Some advantages of the Cox model are that it can utilize many underwritings on the same life and can handle data that is right censored, i.e., subjects can leave the study at any time or the study can end before all subjects have died. The Cox model does not require knowledge of the underlying (base) survival curve, which is advantageous; however, we will see that this advantage also brings challenges when analyzing mortality.

Cox model results are expressed as the logarithm of the hazard so technically, the relative risk factor for each variable is obtained by raising e to the power of the log(hazard). Actuaries will recognize this as consistent with Gompertz. The relative risk factor is interpreted just as it sounds: It describes the force of mortality of subjects having a certain condition relative to that of the reference population, who do not have that condition. A relative risk factor of two for a condition means the subject is twice as likely to die as another subject who does not have that condition.

As an aside, we utilized the survival package in the R statistical language to produce our survival models. It is particularly well-suited for this type of analysis. Other popular statistics programs, such as SAS, also contain survival models using the Cox model.

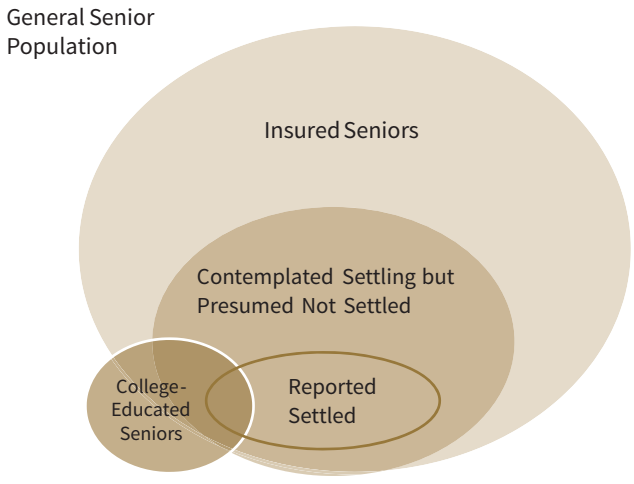
THE ISSUES

A most important issue was that of the underlying mortality distribution. We already had produced mortality tables that varied by age/gender/tobacco use. What then should we do with the Cox model results that also calculated the impact of these variables? It was also very important to ensure that the explanatory variables were truly independent. If not, spurious results would ensue. We also had to redefine certain variables, such as body mass index (BMI), where the risk was actually related to straying from the ideal BMI measurement, rather than the measurement itself. There were many other issues, too numerous to mention in an article of this length.

INPUT DATA

For this exercise, we had available to us over 200,000 underwriting events on 80,000+ unique senior lives, which took place over a 15-year period, primarily in the life settlement market. Figure 1 is a graphic description of the major subpopulations of the universe of senior lives and the populations we studied. At the highest level, there is the general senior population. Some of these seniors have purchased insurance, creating a subpopulation, which can be further broken into two subpopulations:

Figure 1: Senior Populations



> The most important conclusion we drew from this exercise was that despite our best efforts to quantify every aspect of underwriting, there is still considerable judgment brought to bear in the process.

those who actually sold their policies on the secondary market and those who contemplated such a sale but, for some reason, did not conclude the sale. These latter two subpopulations were the basis for our study of antiselection. There is also a small population of college-educated seniors, some of whom can also be associated with the other populations above, which formed the basis for our study of the wealth effect. This data included demographic information such as age, gender, date of birth and date of death. It also included various underwriting conditions such as BMI, smoking status and indicators for various diseases. Included were favorable conditions, such as family history of longevity (parents/siblings who lived beyond age 85) and good exercise tolerance.

CREATING COX PROPORTIONAL HAZARDS MODELS

There was significant data preparation involved. We set up the reference population, which we chose to be males who were age-appropriately active, who did not sell their policies and did not use tobacco. Variables were determined to be either continuous (age, BMI), where the condition has infinite possible values, or binary (coronary artery disease, osteoporosis), where the condition either exists or does not. This required considerable judgment and depended on the availability and form of the data.

Once the data were prepared, we began the process of determining which conditions were statistically significant in predicting mortality. We underwent an iterative process. The Cox models were run with every variable included at first. Then we reran the models, first eliminating most of those variables with a p-value greater than 0.2. This means we were excluding those conditions where the probability that the relative risk shown was due to random fluctuation was over 20 percent. These models were again rerun, this time eliminating those conditions with a p-value greater than 0.1. Finally, we reran the models, including only those conditions where the p-value was at most 0.05.

RESULTS

Figure 2 represents only a portion of the output from our models, consisting of conditions that were included in all runs even if they did not meet the criteria for continued inclusion above. As we advanced through the process, we felt strongly these were fundamental variables that clearly impacted survival and should be included in the analysis regardless of their p-values. In reality, only one variable (rare smoker) would have been eliminated, presumably due to data scarcity. There were a number of other explanatory variables that also made the final cut, but space does not allow their inclusion herein.

Pink/green shading indicates that a condition is hazardous/protective, with the 95 percent confidence limits and p-values also shown. For example, the female hazard is 0.694 of that of males (1.0, as males are the reference). Therefore, the female mortality rate is found by multiplying the male rate by 0.694 for all ages. The hazard for age is 1.08, which means that for any age, the mortality rate for the next higher age is found by multiplying the mortality rate of the first age by 1.08. The smoker hazard is 1.887 times that of the reference, which is nonsmokers; it follows that the smoker mortality rate then is 1.887 times the corresponding nonsmoker rate. This is where the disadvantages of the Cox model came into play. The issue became whether we should replace our base tables for male/female, smoker/nonsmoker with tables based only on the proportional hazards produced in our predictive models and our base male nonsmoker table. After reviewing the model results for consistency with them, we decided to use all four of our existing base tables; however, we broke out antiselection explicitly.

Figure 2

Figure 2	All (<=0.05)				
	Log(hazard)	Hazard	Lower CI	Upper CI	P-Value
Age	0.077	1.080	1.075	1.085	-
Actual BMI less ideal BMI	0.002	1.002	1.001	1.002	0.000
Recurrent Cancer	0.458	1.581	1.365	1.832	0.000
Female	(0.365)	0.694	0.649	0.742	-
Active for their age	(0.141)	0.869	0.802	0.942	0.001
Sedentary	0.200	1.221	1.054	1.415	0.008
Unknown activity level	0.102	1.107	1.031	1.189	0.005
Family history of longevity	(0.087)	0.917	0.857	0.981	0.012
Family history of super longevity	(0.240)	0.787	0.722	0.857	0.000
College-educated population member	0.267	1.306	1.117	1.526	0.001
Settled population member	(0.370)	0.691	0.650	0.734	-
Current smoker	0.635	1.887	1.693	2.103	-
Discontinued smoking	0.178	1.195	1.128	1.267	0.000
Rare smoker	(0.339)	0.713	0.266	1.911	0.501
Tobacco replacement	0.576	1.780	1.187	2.668	0.005
Unknown tobacco use	0.119	1.127	1.018	1.247	0.021

Reference: Male, nonsmoker, normal activity level

CONCLUSIONS

The most important conclusion we drew from this exercise was that despite our best efforts to quantify every aspect of underwriting, there is still considerable judgment brought to bear in the process. However, there is also much useful information that predictive models can provide us because of their ability to process large amounts of data quickly and efficiently. We did validate the antiselection that occurs between those who actually sell their policy versus those who do not (as seen by the hazard ratio of 0.691 for the settled population members in Figure 2). Some results confirmed our clinical judgment; for example, an active lifestyle or family history of longevity are indicators of higher survival rates. Other things went against our clinical judgment; for example, cardiac-related conditions, while still hazardous, were no longer as significant as we thought.

Then there were the confounding results. Hyperlipidemia (high cholesterol) was shown to be protective. We attributed this to the ubiquity of statins. There were a number of other conditions shown to be mildly protective, things such as benign prostatic hyperplasia (BPH), sleep apnea, use of blood thinners and benign colon polyps. We concluded that these were indicators of frequent/better quality of health care, which would allow for early detection and mitigation of more serious risks. Similarly, family history of heart disease and cancer were seen as mildly protective, presumably due to their providing early warning signals to take protective actions, such as better diet and more exercise in the case of heart disease and more frequent screenings in the case of cancers.

BUSINESS OUTCOMES

This analysis was the basis for changes in our debit/credit underwriting model. We replaced an additive model based only on clinical judgment with one that was exponential in nature, which provided more consistency to mortality research. The new model was quite flexible and allowed us to continue to factor in clinical judgment where appropriate. For example, we used the relative risk factor for smokers who quit, but isolated the impact by time since smoking ceased, reducing the debit as time went on. ■

Vincent J. Granieri, FSA, EA, MAAA, is chief executive officer at Predictive Resources LLC. He can be reached at *vgranieri@predictiveresources.com.*

2036: AN ACTUARIAL ODYSSEY WITH AI

By Dodzi Attimu and Bryon Robidoux

Machines currently do what once required human expertise, including tax preparation (United States and other countries), journalism (writing articles based on events), surgery, driving cars, flying aircraft (auto-pilot) and writing software code (e.g., MS Excel macro recorder). We believe most readers would want to know whether machines (software) would someday take away actuarial jobs. In other words, will we be replaced by HAL 9001?[1] Ultimately, a lot, if not all, of what actuaries *currently* do will be taken over by machines in the future. The uncertainty involves the time frames over which the various stages in the transition would take place.

Historically, there have been key phases in interaction of technology with the professions. Starting with the so-called industrial revolution, marked by the advent and use of engines, the next phase was marked by the invention and application of electricity, and the third phase marked by the Internet/web technology explosion. Many observers see us on the verge of a fourth phase, which is an explosion in the use of artificial intelligence (AI) applications.

Many would agree that the past three phases have ultimately led to progress/prosperity for humanity as a whole. However, it remains to be seen if this fourth wave spearheaded by AI is a net positive or negative. In the short term, one thesis is that, so long as the changes are gradual, enabling adaptation by humans, the potential for negative impact will be minimal.[2] Consequently, this would suggest a cause for concern about a potential pending "AI-calypse"[3] if the rate of change is deemed too drastic. There are a number of factors that interact to determine the effects of significant use of AI in the workplace. One of the most important will be the rate of adaptation by the workforce to create value in addition to or to complement what our ever-capable machines do. This topic in the generic sense has been discussed by many authors and along different dimensions, including impact on employment levels, the related issue of resulting distribution of wealth, and ethical issues bound to arise in certain situations.[4] In this article, we consider the potential impact of this fourth wave of technology on the actuarial profession.[5]

MEET THE ROBO ACTUARY AND THE ROBO ACTUARIAL ANALYST

First, let's take a generalized view of the tasks and processes actuaries perform. Actuaries traditionally create and price products that are based on insurable risks and customer demand. This is done by taking into account demographic, economic, regulatory and other external factors. Once the sale is made and the policy is on the books, actuaries set economic and policyholder assumptions for analyzing and managing the product. An important activity is to verify that the business is meeting expectations within many different internal and external metrics and reporting results up the chain, to facilitate decision-making. Generally, the actuary's work involves activities including setting assumptions, building models, analyzing and communicating results, and developing appropriate value-enhancing strategies.

Second, we explain what we mean by the term "robo actuary."[6] A robo actuary is software that can perform the role of an actuary. Though many actuaries would agree certain tasks can and should be automated, we are talking about more than that here. We mean a software system that can more or less autonomously perform the following activities: develop products, set assumptions, build models based on product and general risk specifications, develop and recommend investment and hedging strategies, generate memos to senior management, etc.

Finally, we introduce a closely related term, "robo actuarial analyst," a system that has limited cognitive abilities but can undertake specialized activities, e.g., perform the heavy lifting in model building (once the specification/configuration is created), perform portfolio optimization, generate reports including narratives (e.g., memos) based on data analysis, etc. When it comes to introducing AI to the actuarial profession, we believe the robo actuarial analyst would constitute the first wave and the robo actuary the second wave, which we speculate are achievable in the next five to 10 years and 15 to 20 years, respectively.

WHAT IS AI? WHAT IS ITS CURRENT STATE?

Currently, AI is a buzz word used to lump together different computer science and statistics concepts. At the heart of AI is making intelligent machines that can understand their environment and react accordingly. From the 1950s when John McCarthy coined the term, it is noted in Kaplan (2015), the original goal was to discover the fundamental nature of intelligence[7]and reproduce it electronically. This goal has not been achieved (yet) but progress is being made and many believe it is achievable though the best approach to get there is not unanimously agreed upon. For this article, we will broadly classify AI systems similar to Hawkins and Dubinsky (2016),[7] as belonging to the categories of rule/knowledge-based systems, machine learning systems and "machine intelligence." The latter is based on the core mechanism for exhibiting intelligence.

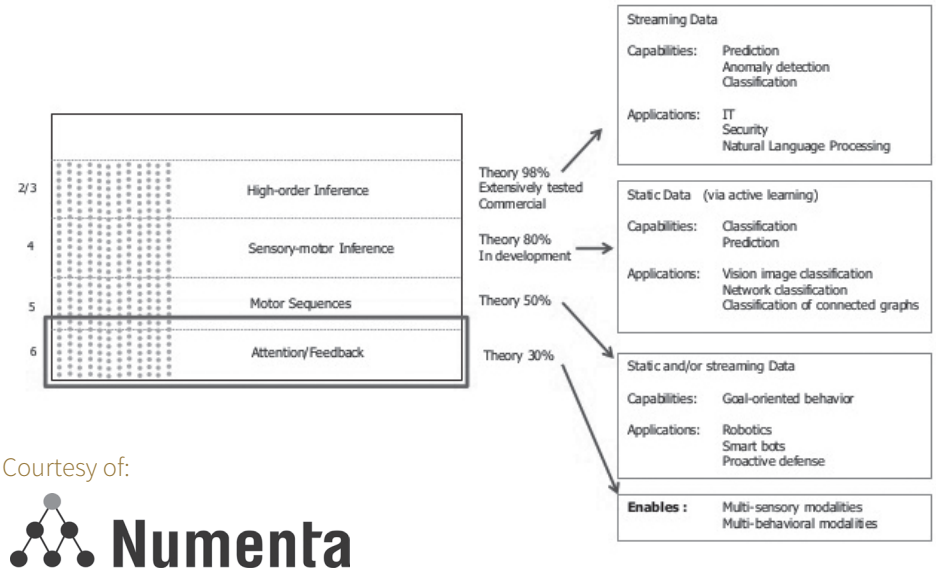
Rule/knowledge-based systems use preprogrammed algorithms and/or look up information to exhibit intelligence. IBM's Watson is a good example of this, as is RGA's AURA e-Underwriting Solution. Classic AI has solved some clearly well-defined problems but is limited by its inability to learn on its own and by the need to create specific solutions to individual problems. In this regard, in spite of it being called artificial intelligence, it has very little in common with general human intelligence.

Machine learning techniques were designed because the rule-based systems become very cumbersome and difficult to maintain/extend. This is because rule-based systems are very problem-specific and any new capabilities have to be laboriously coded and integrated with existing code. Machine learning is a mechanism to find patterns in data without requiring explicit rules. A subset of machine learning is artificial neural networks (ANN), which are based upon 1950s and 1960s understanding of how neural networks work in the brain. ANN methods have evolved into deep-learning techniques. These techniques have splintered off from the original goal of AI to develop machines with brain-like features and focus more on what "works" in a given setting. Deep learning has been able to solve many classification problems, but it needs lots of training data and it can only find static patterns. It fails at recognizing patterns that change and evolve.[8]

Machine intelligence is an approach that seeks to achieve the original goal of replicating human intelligence in electronic form. This approach would have characteristics including the adaptability to different problem domains compared to the general tailored solutions that the first two approaches entail. In that regard, the concept of hierarchical temporal memory (HTM), developed by Numenta, Inc., is probably the most popular attempt toward (true) machine intelligence. HTM is based on the latest research of the neocortex. It simulates how the brain learns in a universal and a continuous way, with robustness to noisy data inputs. One important advantage of HTM over machine learning and classic AI is that the models do not have to be trained manually and there are no tuning parameters. Just like the brain's cognitive processes, HTM is a general purpose problem-solving algorithm. This means the construction of predictive models can be automated. This is the holy grail of AI, because there are massive amounts of data and nowhere near enough data scientists to model it. Numenta has developed an open source project called Numenta Platform for Intelligent Computing (NuPIC), which can be used to develop HTM applications. The following graphic from Numenta shows the current state of the research and what has been commercially developed. The table (courtesy of Numenta) below describes the current understanding of the theory in terms of the four layers of the neocortex.

Finally, we believe many readers can wrap their minds around manual labor or repeatable office activities being taken over by machines, e.g., vacuuming or cleaning the floor, assembly of cars, generation of email alerts, etc. However, with advances in AI in general, and machine learning in particular, computers are proving capable in more and more areas hitherto thought to be limited to the domain of human cognition. For example, machines currently do things like medical diagnosis, surgery, journalism (writing of articles)[9] and driving cars. In fact, there is even a credible expectation in some quarters that artificial intelligent agents will be on major company boards by 2026![10]

Figure 1: Research Roadmap



Courtesy of:

Numenta

ROBO ACTUARIAL ANALYST AND ROBO ACTUARY WITHIN THE AI FRAMEWORK

A robo actuarial analyst is somewhat akin to an actuarial student. It would get tasks with directions from a superior in the organization and may be better/more efficient at specialized tasks than their superiors. In the shorter term, we foresee these systems interacting with human actuaries. In other words, actuaries would perform most of the higher level cognitive tasks to synthesize the lower level heavy-lifting that would then be undertaken by the robo actuarial analyst. This is similar to how one would configure a model to solve an optimization problem. The difference here is the

robo actuarial analyst would be capable of much more than we currently use software systems for. In the next section, we provide an example of how such a robo analyst can do much more than current systems as configured are able to do.

In addition, it is well known that current commercialized AI solutions are more adept/effective in solving specialized problems, e.g., surgery, speech recognition, driving, flying, than general activities like autonomously setting assumptions and making judgments and predictions in a broader-based context, relying on sometimes vague and noisy data. Longer term, machines would be able to handle higher level cognitive actuarial tasks, leading to a scenario where nonhuman systems would interact with the robo actuarial analyst in ways that only human actuaries are able to in the shorter term. This leads us to the concept of a robo actuary. A robo actuary is a system that would have higher cognitive functionality relative to a robo actuarial analyst. We note that from a software architecture perspective, robo actuary and robo actuarial analyst systems could be different components of a single system. We would refer to these systems generally as robo actuarial systems.

A HYPOTHETICAL WORK DAY FOR A ROBO ACTUARIAL ANALYST

We believe most of the heavy lifting work actuaries currently do can be effectively automated, and that is indeed happening. In addition, most of the required underlying technology and framework illustrated in this section is already available.

The robo actuarial analyst will need to be fed the right data/input to perform its processes. One way of simplifying the process for the actuary would be to create a natural language interface that would be higher level than most currently available domain specific languages (DSLs)[11] for the given area of actuarial work. For example, on a given day, a hedging robo actuarial analyst could have an email interface with which an asset-liability management (ALM) actuary could request specific analyses of current hedge positions on the books. A simulation-based analysis would be made with results summarized using both graphics and narrative. The results could be returned with appropriate documents or with links to a central repository of such documents.

Some of the key components of such a system in the light of current technology would include the ability to:

- Map natural language to a set-up/configuration of a simulation model. The building blocks (however rudimentary) of this are already in place, e.g., natural language processing (NLP) solutions including automated voice services on the phone. Taking this a step further, with an appropriate machine-learning capability added to such a system, it should be possible for a component of the system to convert narrative specifying assets/liability characteristics, assumptions and other inputs to create an "internal model representation," which would then be used by the system to generate the software code to create new asset/liability models or update existing ones.

- Run simulation of a hedge strategy. A classical AI system with simulation logic of hedge positions would suffice.

- Generate graphics and narratives from data. Arguably, the leading commercial provider of these services is the firm Narrative Science (see, for example, CITO Research 2016) and their software has been used by firms including financial and news organizations such as Credit Suisse, Nuveen Investments, USAA, CNN and Forbes.

A HYPOTHETICAL WORK DAY FOR A ROBO ACTUARY

As mentioned earlier, the robo actuary would possess higher cognitive skills compared to the robo actuarial analyst. A system that exhibits machine intelligence would possess higher levels of cognition and hence functionality, including dispatching sub-

problems to the more specialized robo actuarial analyst systems because it is predicated on the general purpose problem-solving capabilities of the brain.

Using the NuPIC technology, for example, any work that requires monitoring of trends and analysis of patterns is ripe for automation once neocortex layers 2/3, 4 and 5 are commercially available. Layer 2/3 is currently available but it learns from streaming data, such as market data.[12] Layer 4 is currently in development, but it only learns from slow moving or static data, such as policyholder data. Layer 5 is specialized in goal-oriented tasks, which allows for optimization of profit and capital management from the anomalies and patterns learned from layers 2/3 and 4. Once the policy data is tracked in the admin system and data is streamed from a service such as Bloomberg, the systems can encode the data into a sparse distributed representation (SDR), the data structure utilized by the brain. The SDR allows many different problems to be solved in a uniform way using the HTM algorithm of all the neocortex layers. The SDR is like a computer word with 0 and 1 bits, but, unlike the computer, each bit has a semantic meaning. Given the SDR size is large enough, on the magnitude of 10,000 bits or more, vast amounts of information can be encoded to learn complex patterns, have early detection of anomalies and potentially make goal-oriented decisions.

WILL THIS BE THE END OF THE ACTUARIAL PROFESSION?

To the question of whether AI would mean the end of the actuarial profession, we believe that is not necessarily the case. On the other hand, the profession as we know it today will most likely end.

We believe the increasing use of AI will open other avenues for actuaries. It is conceivable that regulators would still be involved with the various actuarial activities. For instance, regulators are moving away from formula-based reserves to principle-based reserves. This entails moving away from the easy generalization that formulas provided them for regulation. They are concerned with reserve and capital levels in so much as these provide signals of a viable company able to meet its promises to policyholders. HTM-based models could be employed by regulators to determine patterns of healthy companies and provide early anomaly detection to identify failing ones. The beauty of HTM is that it can recognize patterns that change and evolve based on a wide range of metrics without parameter tuning. This will allow regulators to regain the generalization they lost by switching from formula reserves to principle-based reserves. This will enable auditors to focus on more relevant details instead of irrelevant minutiae. Thus, this would present opportunities for actuaries in regulatory, or even auditing roles, to determine principles and standards of practice that are abreast with the times.

In addition, both machine learning and machine intelligence systems rely on the concept of "learning" in that they need to build a representation of the world based on their prior interaction with the world. A key component in the evolution of the robo actuarial systems would be mechanisms of training these systems. We foresee the possibility of an industry for creating solutions that will train these systems to perform their actuarial roles. The SOA and other actuarial bodies will have a part in developing mechanisms to test these systems to ensure they adhere to whatever principles are deemed to be necessary for the health of the actuarial industry and society. Recently, Microsoft launched an AI version of a "teenage girl" called Tay that was supposed to interact with humans on Twitter. From all indications, it seemed like the "training" provided the system did not impose any principles/code of communication, leading to embarrassing tweets from Tay.[13] In a sense, Tay did what it was supposed to do if that was simply to be able to learn how to interact based on tweets it received. It received a good dose of embarrassing tweets and was a quick study to emulate that line of tweeting! In a similar manner, it is possible for robo actuarial systems to learn the wrong things if not properly trained to put their activities within a framework of sorts.

CONCLUSION

In conclusion, though there is concern that machines will take over human work, we believe there is the opportunity for humans to reinvent themselves to be relevant in the light of new developments in artificial intelligence. In particular, we believe the actuarial profession is not exempt from changes due to the increasing involvement of AI in the workplace. On the bright side, with increasing sophistication and intelligence, hopefully HAL 9001 will not have a reason to dominate or even kill us! ■

Dodzi Attimu, FSA, CERA, CFA, MAAA, Ph.D., is director and actuary at Prudential Financial Inc. He can be contacted at *dodzi.attimu@prudential.com.*

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

REFERENCES

CITO Research. 2016. "The Automated Analyst: Transforming Data into Stories with Advanced Natural Language Generation." White paper sponsored by Narrative Science. *https://www.narrativescience.com/automated-analyst?utm_source=NS_Homepage&utm_medium=Uberflip_Pageblock&utm_campaign=Automated_Analyst.*

Colvin, Goeff. 2015. *Humans are Underrated: What High Achievers Know That Brilliant Machines Never Will.* New York: Penguin Publishing Group.

Downes, Larry. 2009. *The Laws of Disruption: Harnessing the New Forces That Govern Life and Business in the Digital Age.* New York: Basic Books.

Egan, Matt. 2015. "Robo Advisors: The Next Big Thing in Investing." *CNN Money,* June 18. *http://money.cnn.com/2015/06/18/investing/robo-advisor-millennials-wealthfront/.*

Floyd, David. 2016. "Can a Robot Do Your Job?" *Investopedia,* Jan. 20. *http://www.investopedia.com/articles/investing/012016/can-robot-do-your-job.asp.*

Hawkins, Jeff, and Donna Dubinsky. 2016. "What is Machine Intelligence vs. Machine Learning vs. Deep Learning vs. Artificial Intelligence (AI)?" Numenta blog, Jan. 11. *http://numenta.com/blog/machine-intelligence-machine-learning-deep-learning-artificial-intelligence.html.*

Hawkins, Jeff, and Sandra Blakeslee. 2005. *On Intelligence.* New York: Times Books.

Hornik, Kurt, Maxwell Stinchcombe and Halbert White. 1989. "Multilayer Feedforward Networks are Universal Approximators." *Neural Networks* 2 (5): 359–66.

Kaplan, Jerry. 2015. *Humans Need Not Apply: A Guide to Wealth and Work in the Age of Artificial Intelligence.* New Haven: Yale University Press.

Leetaru, Kalev. 2016. "How Twitter Corrupted Microsoft's Tay: A Crash Course in the Dangers of AI in the Real World." *Forbes.com,* March 24. *http://www.forbes.com/sites/kalevleetaru/2016/03/24/how-twitter-corrupted-microsofts-tay-a-crash-course-in-the-dangers-of-ai-in-the-real-world/#7321e3e132cb.*

Poole, David, and Alan Mackworth. 2010. *Artificial Intelligence: Foundation of Computational Agents.* New York: Cambridge University Press. *http://artint.info/html/ArtInt.html.*

Susskind, Richard E., and Daniel Susskind. 2015. *The Future of the Professions: How Technology Will Transform the Work of Human Experts.* Oxford: Oxford University Press.

ENDNOTES

1 Readers not familiar with HAL 9001 can refer to http://www.mariowiki.com/HAL_9001.

2 The case is made in Kaplan (2015). The reason there isn't too much pain for the worker for the past few phases is the gradual nature of the change.

3 A play on AI and apocalypse.

4 See Kaplan (2015). A relatively recent fall-out from machines in finance occurred on May 6, 2010 in the stock market sell-off due to algorithmic trading platforms.

5 In Susskind and Susskind (2015), impact of AI on the professions is studied though the actuarial profession is not explicitly included. Professions such as accounting, architecture, medicine, etc., were mentioned.

6 The inspiration for the name comes from the reference to "robo (financial) advisors," e.g., see Egan (2015).

7 Rule-based systems would be equivalent to what Hawkins and Dubinsky (2016) identifies as "classic AI." Also, another common classification could be to consider rule-based and machine-learning systems as "weak AI" and machine intelligence as "strong AI."(See for example, Susskind and Susskind (2015))

8 In Hornik, Stinchcombe and White (1989), where it is shown that any deterministic function of stochastic variables can be approximated by a single-layer neural-network, for example, the author's make it clear that if the functional relationship is stochastic, the results wouldn't hold.

9 As noted in Susskind and Susskind (2015), machine written articles have appeared in reputable information sources as *Forbes*, *Time*, etc. The reader can visit Narrative Science website, *https://www.narrativescience.com/*, for more.

10 Result of a survey of attendees of the 2016 World Economic forum (See for example *http://2serpent.com/2016/01/23/predictions-at-the-2016-world-economic-forum-in-davos-switzerland/*).

11 A DSL is a language for a specific domain, e.g., SQL is a DSL for interacting with relational database systems.

12 An app that does this can be downloaded at *http://numenta.com/htm-for-stocks/.*

13 See Leetaru, Kalev (2016).

14 For example see Floyd, David (2016).

Beyond Multiple Regression

By Michael Niemerg

Suppose you have a large dataset with many independent variables and you want to create a predictive model with only the most significant independent variables. One of the most commonplace approaches in statistics is to apply multiple regression. However, for a dataset with many variables, there is a class of models called penalized regression (aka shrinkage or regularization methods) and least angle regression (LARS) that offer a useful and potentially better alternative to "regular" regression.

To explain these alternate varieties, we need to first backtrack and review simple and multiple regression.

At a cursory level, simple linear regression involves fitting lines to a dataset in a way that minimizes the residual sum of squares (RSS)—more on this later. Most of us probably remember the formula y = mx + b, the "slope intercept" equation of a line. In simple linear regression, y is the variable we are interested in predicting (the response or dependent variable), m is the slope of the line (in regression, these are the coefficients) and b is the y-intercept (β_0 in regression).

At a cursory level, simple linear regression involves fitting lines to a dataset in a way that minimizes the residual sum of squares. ...

The concepts of linear regression can be expanded to contain more than one independent variable (x's). For datasets with potentially many predictive variables, multiple linear regression (and its more sophisticated cousins) is much more manageable, sound and practical than trying to work with independent variables one at a time. To put some notation around it, in multiple regression, we are trying to create a model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_k x_k + \varepsilon$$

In this formulation, y = dependent variable, $x_1, x_2, \ldots x_k$ = the independent variables, β_0 = y-intercept, β_1 = regression coefficients, and ε = random error.

Now, let's motivate the need for alternate forms of regression. One of the difficulties in multiple linear regression is that if a variable is included in the modeling process, a nonzero regression coefficient is generated. This can result in several problems, including overfitting or including statistically significant variables whose effects are small. While there are variable selection methods such as forward selection and backward selection that can help whittle down the list of potential independent variables, they have limitations as well, including high variability and low prediction accuracy when there are many independent variables.

This is where penalized regression comes in. This class of models is good at whittling down a set of potentially many independent variables into something more manageable. It works well when the number of independent variables is large relative to the number of observations. Two other advantages of these models are that they avoid overfitting and their solutions are readily deployable.

In multiple regression, we estimate regression coefficients by minimizing the residual sum of squares. RSS is simply the sum of the squared difference between the actual and predicted response (y).

Equation 1: Quantity Minimized in Multiple Regression

$$RSS = \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

In the formula above, n is the number of observations and p is the number of candidate predictors. Now let's look at the quantity that gets minimized in two of the most common types of penalized regression: least absolute shrinkage and selection operator (LASSO) and ridge to get us an intuitive sense of how they differ.

Equation 2: Quantity Minimized in Ridge Regression

$$RSS\ with\ Penalty\ Term = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda\sum_{j=1}^{p}\beta_j^2$$

Equation 3: Quantity Minimized in LASSO Regression

$$RSS\ with\ Penalty\ Term = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

In the formulas above, the y_i's are the observations, the y_{-i}'s are the predicted values, λ is the tuning parameter and β_j's are the regression coefficients (the parameters we are ultimately trying to estimate).

Notice that extra term on the end in LASSO and ridge regressions? That's where all the magic is. It adds a penalty in the regression formula that places constraints on the size of the regression coefficients. For instance, in LASSO regression, the penalty is the addition of the sum of the absolute values of the regression coefficients multiplied by the tuning parameter. In essence, this penalty shrinks the regression coefficient estimates toward zero to ultimately make them smaller values in the model.

So why do we append this constraint to the equation? Well, it turns out that while adding this tuning parameter adds bias to the regression coefficient estimates, it decreases variability, thereby improving overall prediction error. Another way of thinking about it is that this penalty term prevents us from overfitting our model to our specific data while still allowing us to still find the signal in the noise.

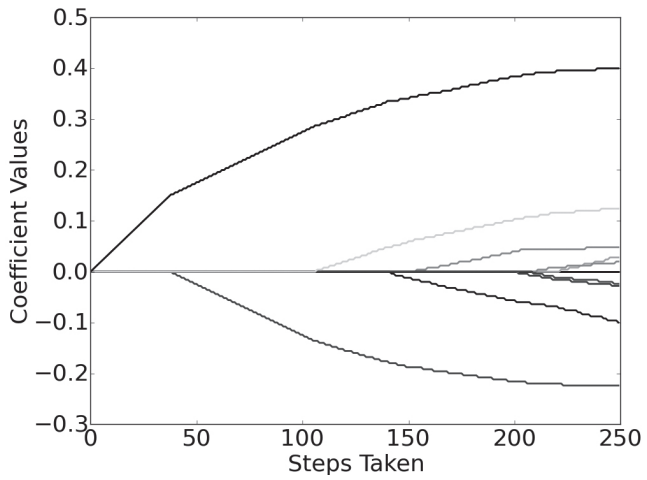
Now, as an astute reader you may be thinking: "That's all well and good but how do we know what value of λ for our tuning parameter to use?" The answer is we don't know, at least not a priori. Rather, we determine the optimal value of λ using cross-validation. That is, we don't train our model on all the data available. Instead, we hold some back to use for testing later. In our initial stage of model building, we only train our model on a subset of the data using multiple values of λ. We then ultimately choose the optimal value λ based on the value that performs best on the data we withheld (there are multiple ways to define "best" here. One way would be to simply use the one that minimizes RSS).

Let's take a look at another methodology related to LASSO and ridge regression called least angle regression (LARS). In LARS, we break the process of fitting the regression coefficient into many small, piecewise steps. In the first step, we start with all the regression coefficients (β_j's) equal to zero. We then find the independent variable that has the highest absolute correlation with the response variable (y) (recall that correlation can range from –1 to 1). We then add a slight increment to this variable's regression coefficient in the direction of its correlation with y. What we have now is a model with one very small nonzero coefficient with all the remaining regression coefficients equal to zero. At this point, we calculate the residuals based on the model we have developed so far and figure out which independent variable has the highest correlation with the residuals and then increment it slightly (it is likely this could be the same predictor for multiple

iterations). We repeat this process iteratively until we reach a predetermined stopping point (for instance, we could decide to take 500 steps, each time incrementing one of the β_j's by .05).

A visualization might help here.

Figure 1: LARS Solution Path



As you can see in Figure 1, different variables are entering the equation at each step. For the first 100 steps in this model, there are only two variables with nonzero coefficients and, as you can see, the value of the coefficient changes with the number of steps (eventually they will plateau). Note that in this chart, all the independent variables were scaled to have mean 0 and standard deviation 1 so that the coefficients values can be easily compared and visualized for magnitude.

One way to think about LARS is to think about it as moving slowly in the direction of multiple regression, one small step at a time. However, we don't need to climb the entire staircase. Instead, we can stop and get off at any time. To determine the optimal stopping point, we can test the model based at various stopping points and use cross-validation to select the best model just like we did with LASSO and ridge regression for the tuning parameter.

One of the advantages of LARS is that it gives us information about how important each variable is to the model and shows us in stepwise fashion how the solution was derived. This is useful in case we want to test how well the model works (using cross-validation) at different points along the solution path. Another advantage is that it performs well when there are lots of independent variables but relatively few observations.

To summarize, the ridge, LASSO and LARS methods are three tools that can help solve some of the shortcomings of multiple regression. They do this by decreasing variability but at the expense of adding bias to the model. There is a trade-off certainly, but, depending on the problem at hand, it might be well worth it.

The world (of regression models) is large. There are many sophisticated models and methods beyond multiple regression that can be useful to a modeler. LASSO, ridge and LARS are a small part of this larger world and just three of many possible tools you could add to your modeling toolbox. Check them out—you'll be glad you did. ■

Michael Niemerg, FSA, MAAA, is an actuary at Milliman in Chicago. He can be reached at *michael.niemerg@milliman.com.*

An Introduction to Incremental Learning

By Qiang Wu and Dave Snell

Machine learning provides useful tools for predictive analytics. The typical machine learning problem can be described as follows: A system produces a specific output for each given input. The mechanism underlying the system can be described by a function that maps the input to the output. Human beings do not know the mechanism but can observe the inputs and outputs. The goal of a machine learning algorithm is to infer the mechanism by a set of observations collected for the input and output. Mathematically, we use (x_i,y_i) to denote the i-th pair of observation of input and output. If the real mechanism of the system to produce data is described by a function f^*, then the true output is supposed to be $f^*(x_i)$. However, due to systematic noise or measurement error, the observed output y_i satisfies $y_i = f^*(x_i)+\epsilon_i$ where ϵ_i is an unavoidable but hopefully small error term. The goal then, is to learn the function f^* from the n pairs of observations $\{(x_1,y_1),(x_2,y_2),\ldots,(x_n,y_n)\}$.

A machine learning algorithm must first specify a loss function $L(y,f(x))$ to measure the error that will occur when we use $f(x)$ to predict the output y for an unobserved x. We use the term unobserved x to describe new observations outside our training sets. We wish to find a function such that the total loss on all unobserved data is as small as possible. Ideally, for an appropriately designed loss function, f^* is the target function. In this case, if we can compute the total loss on all unobserved data, we can exactly find f^*. Unfortunately, computing the total loss on unobserved data is impossible. A machine learning algorithm usually searches for an approximation of f^* by minimizing the loss on the observed data. This is called the empirical loss. The term **generalization error** measures how well a function having small empirical loss can predict unobserved data.

There are two machine learning paradigms. **Batch learning** refers to machine learning methods that use all the observed data at once. **Incremental learning** (also called online learning) refers to the machine learning methods that apply to streaming data collected over time. These methods are used to update the learned function accordingly when new data come in. Incremental learning mimics the human learning process from experiences. In this article, we will introduce three classical incremental learning algorithms: the stochastic gradient descent for linear regression, perceptron for classification and incremental principal component analysis.

STOCHASTIC GRADIENT DESCENT

In linear regression, $f^*(x) = w^Tx$ is a linear function of the input vector. The usual choice of the loss function is the squared loss $L(y,w^Tx) = (y-w^Tx)^2$. The gradient of L with respect to the weight vector w is given by

$$\nabla_w L = -2(y - w^T x)x.$$

Note the gradient is the direction for the function to increase, so if we want the squared loss to decrease, we need to let the weight vector move opposite to the gradient. This motivates the stochastic gradient descent algorithm for linear regression as follows: the algorithm starts with the initial guess of w as w_0. At time t, we receive the t-th observation x_t and we can predict the output as

$$\widehat{y_t} = w_{t-1}^T x_t.$$

After we observe the true output y_t, we can update the estimate for w by

$$w_t = w_{t-1} + \eta_t(y_t - \widehat{y_t})x_t$$

The number $\eta_t>0$ is called the step size. Theoretical study shows that w_t becomes closer and closer to the true coefficient vector w provided the step size is properly chosen. Typical choice of the step size is

$$\eta_t = \frac{\eta_0}{\sqrt{t}}$$

for some predetermined constant η_0. Another quantity to mea-

sure the effectiveness is the accumulated regret after T steps defined by

$$Regret = \sum_{t=1}^{T}(y_t - \hat{y_t})^2 - \sum_{t=1}^{T}(y_t - w^T x_t)^2$$

If this algorithm is used in a financial decision-making process and $w^T x_t$ is the optimal decision at step t, the regret measures the total **additional**[1] losses because the decisions are not optimal. In theory, the regret is bounded, implying that the average additional loss resulting from one decision is minimal when T is large.

We use a simulation to illustrate the use and the effect of this algorithm. Assume that in a certain business, there are five risk factors. They may either drive up or down the financial losses. The loss is the weighted sum of these factors plus some fluctuation due to noise: $y = x_1 - x_2 + 0.5x_3 - 0.5x_4 + x_5 + \epsilon$. So the true weight coefficients are given by w=[1, -1, 0.5, -0.5, 2]. We assume each risk factor can take values between 0 and 1 and the noise follows a mean zero normal distribution with variance 0.01. The small variance choice is empirically selected to achieve a smaller signal to noise ratio. We generate 1,000 data points sequentially to mimic the data-generating process and perform the learning with an initial estimate w_0=[0,0,0,0,0]. In Figure 1, we plot the distance between w_t and w, showing estimation error decays fast (which is desirable). In Figure 2, we plot the regret for each step. We see most additional losses occur at the beginning because we have used a stupid initial guess. They increase very slowly after 50 steps, indicating the decisions become near optimal. In other words, even a poor guess can lead to excellent results after a sufficient number of steps.

Figure 1: Estimation Error vs. Iterations

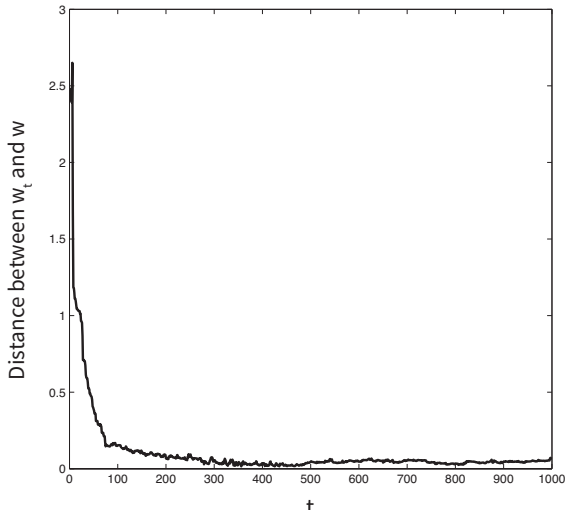
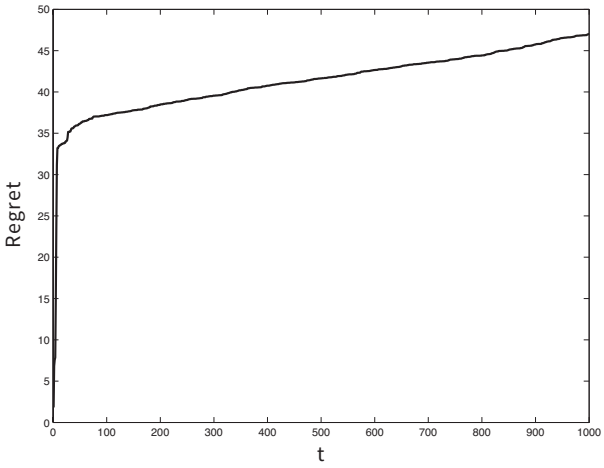


Figure 2: Regret vs. Iterations



PERCEPTRON

In a classification problem, the target is to develop a rule to assign a label to each instance. For example, in auto insurance, a driver could be labeled as a high risk or low risk driver. In financial decision-making, one can determine whether an action should be taken or not. In a binary classification problem where there are two classes, the labels for the two classes are usually taken as 0 and 1 or –1 and +1. When –1 and +1 are used as the two labels, the classifier could be determined by the sign of a real valued function. A linear classifier is the sign of a linear function of predictors $f(x) = sign(w^T x)$. Mathematically $w^T x = 0$ forms a separating hyperplane in the space of predictors. The perceptron for binary classification is an algorithm to incrementally update the weight vectors of the hyperplane after receiving each new instance. It starts with an initial vector w_0 and when each new instance (x_t, y_t) is received, the coefficient vector is updated by

$$w_t = \begin{cases} w_{t-1} + \eta_t y_t w_t, \text{if } y_t(\beta_{t-1} x_t) < \gamma, \\ w_{t-1}, \end{cases}$$

otherwise, where γ is a user specified parameter called the **margin**. The original perceptron introduced by Rosenblatt in the 1950s has a margin 0, i.e., $\gamma = 0$. The perceptron can be explained as follows. If $y_t(\beta_{t-1} x_t) < 0$, the t-th observation is classified incorrectly and thus the rule is updated to decrease the chance for it being classified incorrectly. If $y_t(\beta_{t-1} x_t) > 0$, the t-th observation is classified correctly, and no update is necessary. The idea of using a positive margin is from the well-known support vector machine classification algorithm. The motivation is that the classification is considered unstable if the observation is too close to the decision boundary even when it is classified correctly. Updating is still required in this case as a penalty. The classification rule is not updated only when an instance is classified correctly

Principal component analysis (PCA) is probably the most famous feature extraction tool for analytics professionals.

and has a margin from the decision boundary. For perceptron, the cumulative classification accuracy, which is defined as the percentage of the classified instances, can be used to measure the effectiveness of the algorithm.

In Figure 3, we simulated 1,000 data points for two classes: the positive class contains 500 data points centered at (1, 1) and the negative class contains 500 data points centered at (–1, –1). Both classes are normally distributed. The optimal separating line is $x_1 - x_2 = 0$, which can achieve a classification accuracy of 92.14 percent. That is, there is a systematic error of 7.86 percent. We assume the data points come in sequentially and apply the perceptron algorithm. The cumulative classification accuracy is shown in Figure 4. As desired, the classification ability of the perceptron is near optimal after some number of updates.

Figure 3: Data for a Binary Classification Problem

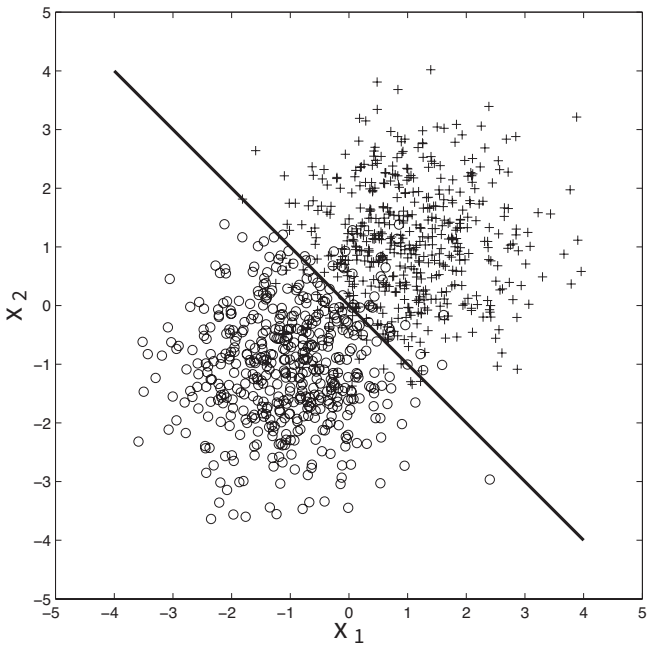
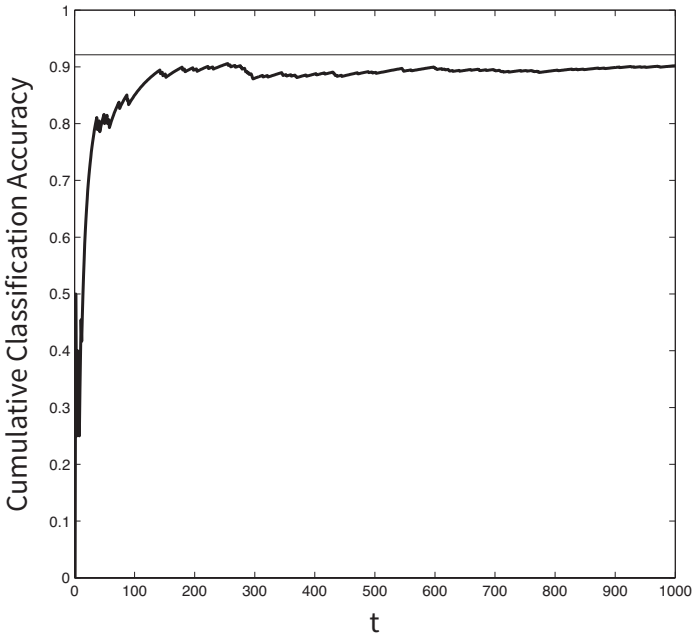


Figure 4: Cumulative Classification Accuracy of Perceptron Technique



INCREMENTAL PCA

Principal component analysis (PCA) is probably the most famous feature extraction tool for analytics professionals. The principal components are linear combinations of predictors that preserve the most variability in the data. Mathematically they are defined as the directions on which the projection of the data has largest variance and can be calculated as the eigenvectors associated with the largest eigenvalues of the covariance matrix. It can also be implemented by an incremental manner. For the first principal component v_1, the algorithm can be described as follows. It starts with an initial estimation $v_{1,0}$ and when a new instance x_t comes in, the estimation is updated by

$$u_{1,t} = (t-1)v_{1,t-1} + \left(v_{1,t-1}^T x_t\right)x_t,$$
$$v_{1,t} = \frac{u_{1,t}}{\|u_{1,t}\|}.$$

The accuracy can be measured by the distance between the estimated principal component and the true one.

Again, we use a simulation to illustrate its use and effectiveness. We generated 1,000 data points from a multivariate normal distribution with mean μ = [1,1,1,1,1] and covariance matrix

$$\begin{bmatrix} 4 & -1 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0.2 & 0 & 0 \\ 0 & 0 & 0 & 0.2 & 0 \\ 0 & 0 & 0 & 0 & 0.2 \end{bmatrix}.$$

The first principal component is [0.9517, –0.2898, 0, 0, 0]. In Figure 5, we used the scatter plot to show the first two variables of the data with the red line indicating the direction of the first principal component. After applying the incremental PCA algorithm, the distance between the estimated principal component and the true principal component is plotted for each step in Figure 6. As expected, the distance shrinks to 0 as more and more data points get in.

Figure 5: Feature Abstraction via Principal Component Analysis

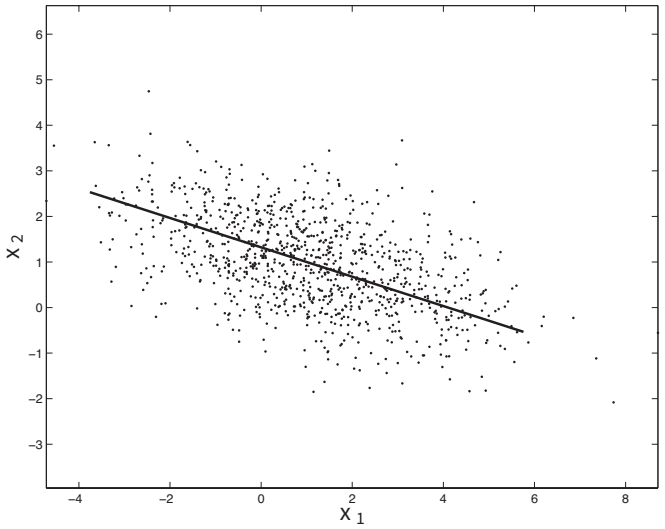
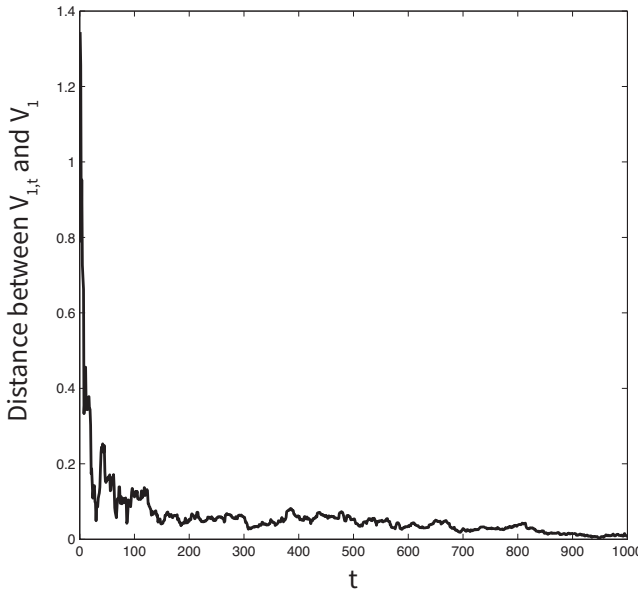


Figure 6: Estimation Error from Principal Component Analysis



REMARKS

We close with a few remarks. First, incremental learning has very important application domains, for example, personalized handwriting recognition for smartphones and sequential decision-making for financial systems. In the real applications, batch learning methods are usually used with a number of experiences to set up the initial estimator. This helps avoid large losses at the beginning. Incremental learning can then be used to refine or "personalize" the estimation. Second, we have introduced the algorithm for linear models. All these algorithms can be extended to nonlinear models by using the so-called kernel trick in machine learning. Finally, we would mention that it seems the term "online learning" is more popular in machine learning literature; however, we prefer the term "incremental learning" because "online learning" is widely used to refer to the learning system via the Internet and can easily confuse people. Actually, in *Google*, you probably cannot get what you want by searching "online learning." Instead, "online machine learning" should be used. ■

Qiang Wu, PhD, ASA, is asociate professor at Middle Tennessee State University in Murfreesboro, Tenn. He can be reached at *qwu@mtsu.edu.*

Dave Snell, ASA, MAAA, is technology evangelist at RGA Reinsurance Company in Chesterfield, MO. He can be reached at *dave@ActuariesAndTechnology.com.*

ENDNOTES

1. Vladimir N. Vapnik, Statistical Learning Theory, John Wiley & Sons,1998.
2. Juyang Weng, Yilu Zhang, and Wey-Shiuan Hwang, Candid Covariance-Free Incremental Principal Component Analysis, IEEE Transactions on Pattern Analysis and Machine Intelligence, 25(8), 2003, 1034-1039.
3. Wikipedia, Online Machine Learning. *https://en.wikipedia.org/wiki/Online_machine_learning*
4. Wikipedia, Perceptron. *https://en.wikipedia.org/wiki/Perceptron*

Follow Your Passion

By Shea Parkes

I got an aggressive start in actuarial science: I evaluated colleges based on their actuarial science departments and did not look back until I achieved my FSA. In my decade at Milliman, I have had at least three different careers. After spending my first years as a bumbling beginner, I was briefly a consultant. Now I productionize data-focused solutions and consider myself somewhat of a statistically focused "intrapreneur" (an entrepreneur who works from within a large organization).

In all of my career phases, my biggest joy has been learning. SOA exams grounded me in the ideas of actuarial credibility and the intricacies of the U.S. health care system, while consulting has helped me focus on solving relevant business problems. Trying to maintain long-term successful solutions showed me the need to know more about software development.

While doing traditional consulting work, I would have the pleasure of assisting on a valuable solution that could be expanded to multiple clients. At first we would just copy and paste everything, and then alter the copy until it worked for the next client. As my colleagues and I got better at problem-solving with applied statistics, our solutions started living longer and longer, and we were eventually maintaining years-old solutions. A common (and true) software idiom is that the worst code you will ever see is the code you wrote six months ago. A few of us had recently finished our actuarial examinations and felt like we had the appetite to learn more and do better.

The software development profession does an excellent job encouraging self-learning; many resources are available. My personal learning style is to consume a torrent of text. I read a mix of current blogs and authoritative textbooks. The textbooks impart a deeper understanding of complex concepts, while the blogs provide a broader picture of modern best practices (and pain points). I would quickly jump between specific subjects as they became important to my current duties; this way, I was always reinforcing what I was reading with applied practice. I feel a large tipping point at this stage in my career was the transition away from spreadsheets and toward fully embracing modern revision control systems (e.g., Git and GitHub) for everything we did.

The ability to produce robust, reusable, extensible, testable, maintainable and automated solutions is invaluable. I am still not an expert (and may never be), but I know a lot more now than an average actuary. I work with professional software developers and can often view problems and solutions from their perspectives. I can work with the tools they use and offer meaningful contributions to the analytics components of our products.

In particular, the solid data intuition gained during my earlier years is a great asset to have in the field of software development. Not only can I quickly dismiss some results as incorrect, I can often give helpful suggestions as to which stage in an analytics pipeline most likely contains the responsible errors.

A pure computer science education can leave somewhat of a blind spot when it comes to the meaning behind data. Data might be treated as an inconsequential abstract quantity, or a black box best left unopened. Software projects often rely upon business analysts writing (excessively detailed) requirements documents to ensure the results are solving the right problems. This is an understandable division of labor, but not one we can afford to utilize for all changes or enhancements, which is due to our team size. All of our developers currently get to dabble in deployment and operations. The actuarial expatriates like myself spend more time directly committing changes to pertinent business logic (and authoring appropriate unit/regression tests) than documenting what needs to be done for others.

We are continuously trying to improve our new hire training, and this has had to be adapted to the hiring of dedicated computer science graduates. We have found it valuable to dedicate time in their first weeks to familiarizing them with common health care data sources. We explain why these sources were collected and how we utilize them. We try hard to ensure they are not just abstract tables and fields.

I have greatly enjoyed learning more about the field of software development. I feel it has shaped my career and my abilities in a positive direction. I do not know how long my career will stay in this phase, but I feel I am bringing valuable context and contributions to my new colleagues. ■

Shea Parkes, FSA, MAAA, is an actuary at Milliman Inc. He can reached at shea.*parkes@milliman.com.*

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.

An Insurance Company for the 21st Century: A Thought Experiment

By Jeff Huddleston and Benjamin Smith

Over the past decade, many industries have been disrupted by companies that have leveraged technology and data in unique and powerful ways. Imagine an insurance company named Brightly Co. has been founded with predictive analytics and big data embedded as its core functions. In this hypothetical scenario, a news reporter has been dispatched to interview Brightly's CEO and to learn more about how the company operates. The article follows.

BRIGHTLY SHINES IN E-INSURANCE FIELD

Eschewing traditional distribution models and distributing exclusively online has shown to be a successful approach for Brightly Co., based in New York. Consumers, used to buying everyday goods from the online behemoth, found the process of buying a simple term life, auto or health insurance policy online easy and familiar.

"We decided early on that the traditional process of buying insurance, such as through a broker, is totally obsolete for certain segments of the business," said Sam Coleman, Brightly's CEO and founder. "It seemed radical at the time, but we decided that based on changes in consumer behavior, it made total sense to sell our products through an e-commerce provider."

Not only has Brightly made the distribution process more consumer friendly by leveraging an e-commerce platform, the company has developed cutting-edge techniques to significantly lessen its underwriting time. For instance, Brightly has developed a process called Expedited Issue to make its underwriting for life and health products best-in-class. Expedited Issue is powered by predictive algorithms and an abundance of data. However, Coleman noted, the emergence of electronic health records have been critical in developing the program.

In 2009, Congress passed the American Recovery and Reinvestment Act which, among other things, established a timeline for future incentives for health care providers to offer patient health records in electronic format. Since then, the robustness of electronic health care data has improved dramatically.

"Previously, to complete the underwriting process, insurers required patients to visit their doctor, have blood drawn and provide a list of previous medications," Coleman said. "Back in 2009, we could envision a world where a large amount of the information gained from such an intrusive and lengthy process would be obtained by simply downloading a patient's health records. Today, that world is a reality, and we have incorporated electronic health records into our underwriting process to gain a competitive advantage."

Indeed, while most of Brightly's competitors still adhere to the traditional methods of underwriting (which are both costly and time-consuming), Expedited Issue allows Brightly to determine which applicants are so clearly qualified for life and health insurance that they can be underwritten within mere seconds.

For an applicant who does not qualify for Expedited Issue, Brightly's predictive algorithms identify which medical underwriting requirements are needed to underwrite an applicant and orders only those pieces. Oftentimes the company only needs one or two data points to determine how an applicant should be underwritten. For a hypothetical example, it is possible that Applicant A was ordered to have blood tests because Brightly's model anticipated a likelihood of Ailment A in his activity, whereas Applicant B was ordered to have an attending physician statement (APS) because the model thought she was likely to develop Ailment B. Being able to order piecemeal requirements based on each applicant's data enables Brightly to avoid costly medical expenses traditionally associated with underwriting and minimizes the invasiveness and lengthiness of the underwriting process for the policyholder.

Coleman summed it up this way: "Typically it takes an insurer anywhere from 30 to 90 days to underwrite an applicant for life

insurance. With Brightly, if you qualify for Expedited Issue, you are approved within minutes. Even if you're not granted Expedited Issue, the process is still more pleasant and much faster than the traditional underwriting experience for life or health insurance."

Because customers who are granted Expedited Issue cost Brightly very little to acquire and are expected to have lower claims, they are highly coveted. As such, the company has developed predictive algorithms to identify consumers in the United States who are not only most likely to buy Brightly's insurance, but are most likely to qualify for Expedited Issue. The customers who score the highest (the most likely to buy and qualify for Expedited Issue) are targeted aggressively through customized marketing based on Brightly's algorithms. For instance, an older applicant might receive a flier in the mail whereas a millennial might see content sponsored by Brightly appear in one of their social media news feeds.

POLICY LIFETIME

Brightly has further differentiated itself through its innovative Flexible Premium Program where policyholders can elect to share personalized data from a variety of sources (such as a policyholder's smart watch, cell phone or Internet-enabled home devices). If the data indicates that a policyholder is exhibiting low-risk behavior, they may qualify for lower premiums that month. This benefits both Brightly and the policyholder: Brightly can more effectively manage risk and policyholders pay lower premiums.

Brightly can utilize data from almost any Internet-enabled device. Such devices offer a trove of lifestyle data that is extremely valuable to a life/health insurer trying to better understand its policyholders' behavior. For instance, many smart devices not only remind users to take walks when they have been sitting for an extended period of time, but will also track the number of steps taken. Devices can also estimate the number of calories consumed by analyzing wireless purchases. Some devices can even track users' vital signs and call emergency services in the event the device detects a high probability of an imminent heart attack or stroke. Obviously, an insurance company would like for its policyholders to utilize many of these features since they help prevent claims. Brightly makes this easy—policyholders simply elect to share data with Brightly through their device's settings center.

Of course, with the rise of the "Internet of Things," the Flexible Premium Program is not just limited to life and health insurance products—a policyholder can elect to share data from their Internet-enabled vehicle or Internet-enabled devices at home. Does a policyholder listen to music at a safe volume when driving, avoid dangerous intersections and obey speed limits? If so, Brightly may discount the policyholder's auto insurance by a few percentage points. Does the policyholder check the Internet-enabled smoke detector in his or her home monthly to make sure it is working? Does the policyholder confirm, via his or her phone, that the doors to their home are locked and the oven is off on a daily basis? They might see a reduction in their homeowner's insurance.

THE LONGITUDINAL EFFECT

In addition to driving down policyholder premiums, Brightly's innovative use of data allows the company to build a broad picture of its policyholders. Indeed, most insurance companies receive information on a policyholder when insurance is purchased and when a claim is made. However, Brightly learns about its policyholders' behaviors and habits throughout the policy's lifetime. In other words, on a policyholder level, the company's data set is not limited to a few discrete data points but a continuous story constructed and refined throughout the policyholder's lifetime.

This wealth of data is a major boon to Brightly.

"We have a better understanding of how our policyholders behave over time," Coleman said. "For instance, there are policyholders who did not smoke when they bought their policies. However, over time, they picked up the habit. Well, we can use that information to refine our predictive algorithms to better underwrite applicants who have similar traits."

... Brightly learns about its policyholders' behaviors and habits throughout the policy's lifetime.

Coleman listed other examples. "Let's say a policyholder lives in New York and has exhibited great health choices and that policyholder gets a new job in Los Angeles. Well, that person is probably going to need a car and, based on their healthy lifestyle choices, we think there is a high probability they will be a good driver. Let's try to cross-sell them auto insurance.

"Other companies try to do these things but really struggle. We excel because data and predictive analytics are natural parts of our operations."

NONTRADITIONAL BENEFITS

"Other insurance companies exist purely to enter a financial agreement with their policyholder. We certainly do that as well, but we think that Brightly, more fundamentally, is an extension

of our policyholders' well-being," Coleman said. "Sure, our insurance products protect our policyholders in an unforeseen event, but Brightly is so much more—we work in harmony with our policyholders to help them live healthier lifestyles."

Indeed, Brightly has partnered with other companies to encourage its policyholders to make healthy decisions. For example, while at a food court, a policyholder might receive a notification that salads at Only Organic Options (a restaurant specializing in locally sourced organic food) are being discounted by 30 percent and that buying one would count toward reducing their monthly life and health premiums. The idea is that even if the policyholder is craving pizza and soda, they will be incentivized to choose a locally sourced Greek salad.

"It's a win-win-win!" Coleman said. "The policyholder has lower premiums, Brightly expects to pay less benefits in the future, and Only Organic Options has more business. Moreover, Brightly brings in additional revenue from Only Organic Options for bringing them business. Oftentimes, it is enough to offset the reduction in premium offered to the policyholder. At the end of the day, however, what we are really doing is helping our policyholders lead healthier lifestyles."

CONCLUSION

Because data and analytics are programmed into Brightly's DNA, the company is well positioned to fuel growth through innovative analytic solutions. Coleman imagines that within the next two years Brightly will introduce One-Day Insurance where policyholders can purchase insurance for a single day if they know they will be participating in risky activities (for example, dangerous skiing or skydiving).

Like any insurance company, we won't know if Brightly has underwritten and priced its products appropriately until more claims experience emerges. However, the company's use of predictive analytics has revolutionized the insurance industry and slashed costs across the insurance business lifecycle. As the amount of data in the world increases, Brightly is well positioned to challenge existing insurance practices and bring new analytically driven innovations to the market. ■

Both Jeff and Ben specialize in helping clients understand how the application of predictive models can improve operations throughout their organization.

Jeff Huddleston, ASA, CERA, MAAA, is a senior consultant at Deloitte Consulting LLP. He can be reached at *jehuddleston@deloitte.com.*

Ben Smith, ASA, MAAA, is a consultant at Deloitte Consulting LLP. He can be reached at *bensmith@deloitte.com.*

Three Pitfalls to Avoid in Predictive Modeling

By Marc Vincelli

Few would disagree with the power and promise of predictive modeling. From the Oakland A's use of predictive modeling to build a championship baseball team on a shoestring budget in 2002, to Google's use of search and text analytics to predict the H1N1 flu outbreak in 2009, the well-known examples of predictive modeling "successes" are numerous. Perhaps less widely recognized is the myriad of ways in which a predictive model can fail to perform as expected, often due to misconceptions or misrepresentations on the part of the analyst. In this article, I focus on three such pitfalls.

1. FORCING A PREDICTIVE MODEL ON A PROBLEM IN THE WORLD OF UNCERTAINTY

Economists and decision theorists have for some time distinguished between decisions made under risk and decisions made under uncertainty. In the world of risk, all alternatives, consequences and probabilities are known, or can be reasonably developed (using past experience, for example). In the world of uncertainty, some of this information is unknown, and possibly even unknowable.[1] While decision problems in the world of risk lend themselves well to statistical thinking, those in the world of uncertainty require good rules of thumb (heuristics)[2] and expert intuition balanced by deliberative reasoning.

The nuanced distinction between risk and uncertainty is important to consider when determining whether the predictive modeling toolkit is even appropriate for a given prediction problem. Some problems, such as predicting long-term interest rates or forecasting an individual's future financial needs, may involve too much uncertainty to appropriately leverage predictive modeling. In these cases, the application of professional judgment informed by a simulated fan of outcomes, in line with the RAND Corporation's robust decision-making (RDM) framework,[3] may be more prudent. Forcing a predictive model on a problem that resides within the world of uncertainty can result in suboptimal business decisions, a false sense of comfort and serious financial consequences. So before going too far down the predictive modeling path, the analyst is well advised to ask himself: "Am I dealing with a prediction problem in the world of risk, or uncertainty?"

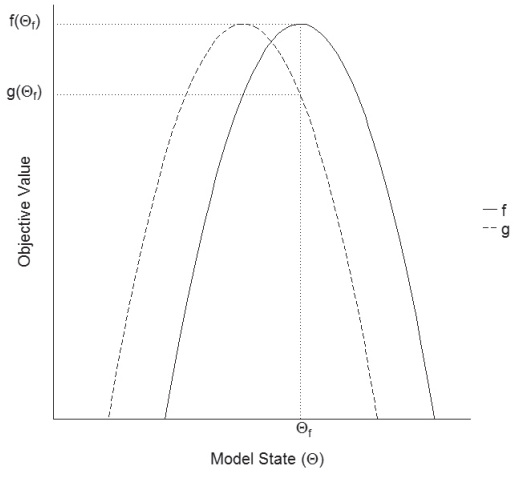
2. SUGGESTING THAT FUTURE MODEL RESULTS ARE LIKELY TO BE AS GOOD AS PAST RESULTS

Any model that has been optimized based on past data is likely to experience performance degradation upon implementation. This phenomenon, in which optimization produces a model that is more likely to perform well in the future but less likely to perform as well as past results suggest, has been called the *optimization paradox*.[4]

To see why future results are not likely to be as good as past results, consider the optimization of the objective function f in Figure 1. An objective function can be thought of as relating a quantity of interest (e.g., profit) to various model states (Θ) based on data available at the time the function was generated. Assume function f is the objective function we obtained just prior to model implementation. Under an optimization approach, we would implement the model state that maximizes (or minimizes, as appropriate) our objective function. Let us denote this optimized model state as Θ_f.

Now assume we have implemented the model and have accumulated more experience. Unless the future is just like the past, we can expect the emerging data to shift f in some unpredictable way, resulting in a new objective function g. How might the objective function shift? Well, in the neighborhood of the optimum (Θ_f , $f(\Theta_f)$), which is our area of interest, the primary shifts would be up, down, left or right. Figure 1 illustrates the left-shifted case. The key to understanding the optimization paradox is to recognize that in three out of the four primary translations (i.e., shift down, shift left and shift right in our example), $g(\Theta_f)$ will be less than $f(\Theta_f)$. In other words, most of the time we should not expect future model results based on Θ_f to be as good as past results.

Figure 1: Illustration of Optimization Paradox with Left-Shifted Objective Function



Does this mean the analyst should avoid optimization? Absolutely not; optimization produces the "best" answer for a given objective function generated at a point in time, and a solution that will more likely than not continue to outperform its suboptimal alternatives. What it does mean, however, is that the analyst must appropriately set model performance expectations with the end-user. One way in which this can be done is to favor out-of-sample test results over in-sample test results when discussing expected performance.

3. OVER-SEARCHING TO FIND PATTERNS AND RELATIONSHIPS

One of the dangers with building predictive models on big data is over-searching, which can lead to spurious correlations and nonsensical models. If we dredge through enough data, we will eventually—by chance alone—find something that appears to be correlated to our target variable but really has no relationship whatsoever. It is incumbent on the analyst to apply his own professional judgment to validate the inclusion of variables and to avoid testing a hypothesis on variable inclusion with evidence used in constructing the hypothesis itself.

Perhaps one of the best known examples involving spurious correlation is the Super Bowl Indicator, which "predicted" that when a premerger National Football League team won the Super Bowl, the U.S. stock market would rise, and when an old AFL team won the Super Bowl, the U.S. stock market would fall. It turns out that between 1967 and 2013, this indicator was correct more than 70 percent of the time. Surprisingly, the indicator was even discussed in the highly respected *Financial Analysts Journal*.[5] So would you be willing to put your money and/or reputation on the line that this correlation is predictive? Only one's good judgment, and not a model, can answer that question.

CONCLUSION

As powerful and promising as predictive modeling can be, practitioners have a responsibility to ensure that the toolkit is applied appropriately and that end-users understand each model's "sphere of competence" (including intended usage, expected performance and risks). Three steps one can take toward this end are to:

- avoid applying predictive modeling to problems that reside within the world of uncertainty,
- explain to the end-user that future model results are unlikely to be as good as results optimized to the training data, and
- identify and exclude variables with spurious correlations. ■

Marc Vincelli, ASA, M.Sc., is a principal consultant with Fortis Analytics in Kitchener-Waterloo, Canada. He can be reached at *marc_vincelli@fortisanalytics.com.*

ENDNOTES

1 Pablo A. Guerron-Quintana, "Risk and Uncertainty," Philadelphia Fed Business Review Q1 (2012): *https://www.phil.frb.org/research-and-data/publications/business-review/2012/q1/brq112_risk-and-uncertainty.pdf.*

2 Gerd Gigerenzer, *Risk Savvy: How to Make Good Decisions* (New York: Penguin Books, 2014).

3 Robert J. Lempert, et al., "Making Good Decisions Without Predictions: Robust Decision Making for Planning Under Deep Uncertainty," *RAND Corporation Research Briefs* RB-9701 (2013): *http://www.rand.org/pubs/research_briefs/RB9701.html.*

4 Curtis M. Faith, *The Way of the Turtle* (New York: McGraw-Hill, 2007).

5 Robert R. Johnson, "Is It Time to Sack the 'Super Bowl Indicator'?" Total Return (blog), *The Wall Street Journal*. Jan. 22, 2014, *http://blogs.wsj.com/totalreturn/2014/01/22/is-it-time-to-sack-the-super-bowl-indicator.*

The Actuarial Road Not Taken: Jeff Sagarin's Sports Ratings

By Anders Larson

To close out the 2015-16 college football season, the Clemson Tigers were set to play the Alabama Crimson Tide. Like many matchups in college football's postseason, the two teams had not played each other during the regular season. In fact, they had not played each other since 2008.

So who should be expected to win? Well, one place to look would be the point spread. The point spread for each game is devised by gambling organizations, and it is designed to handicap the game such that each team should have roughly an equal chance to win, after adding in the point spread.[1] For instance, the point spread between Clemson and Alabama was seven points in favor of Alabama at most sports books, meaning gamblers betting on Alabama would need Alabama to win by at least seven points to win the bet.

Point spreads have been around for years, and ever since their inception, sports gamblers and casual fans alike have looked for ways to outsmart the oddsmakers. Back in the early 1970s, one of those fans was Jeff Sagarin, a recent graduate of the Massachusetts Institute of Technology. At the time, Sagarin was considering a career as an actuary. He passed three actuarial exams in the late 1960s and early 1970s and worked at New York Life as an actuarial trainee for a brief period.

Many gamblers would look to superstition or misguided "statistics" (hey, the Yankees have won their last seven Tuesday games against a right-handed pitcher—they're a lock![2]). Sagarin took a more analytical approach: He decided to devise a rating system that would help predict both the outcome and the margin of victory if two teams played each other. His ratings differed from the traditional poll rankings, which often came down to a subjective opinion of which teams were most "deserving."

"I never even thought about the 'reward' thing," Sagarin said. "I wanted to predict games as accurately as possible. I wanted to see if I could be as accurate as the point spreads in the New York Post."

Sagarin's system was data-driven, taking into account scores from games across the country (and other variables, such as home-field advantage). Aggregating all the scores was critical, but at the time, simply collecting those scores was often the hardest part of the process. As actuaries, we understand the challenge of aggregating and cleaning data and how critical it is to our work product. But most of us who entered the industry in the past 20 years probably haven't had to go to the same lengths that Sagarin did back in the 1970s. On a typical day, Sagarin, living in Boston, would drive to the office of the Boston Globe and cut out its wire with each day's basketball scores. But since the Globe's wire was often incomplete, Sagarin had to resort to plan B.

"So sometimes what he'd have to do was, say there was a small school, say Ball State in Indiana," said Larry Isaacs, a long-time friend of Sagarin's. "What he'd do is call telephone information in Muncie, Indiana, and get the operator on the line, and he'd sweet talk the operator because he was pretty charming. And he'd say, 'By the way, was there a basketball game last night?' And he'd say, 'Can you tell me who won that game?' That's how we'd get the scores. It was really low-tech in those days."

After finding he was having some success, Sagarin caught on with some magazines and newspapers, including Pro Football Weekly and the Boston Globe. He ultimately decided to forego actuarial science and make a full-time career out of his sports rating systems. His big break came in 1985, when his ratings started appearing in USA Today, where they still appear for several sports. Sagarin's ratings, which have been around far longer than most other rating systems currently on the market, consistently have among the most accurate predictions. ThePredictionTracker.com evaluates nearly 70 college football rating systems.[3]

Sagarin's primary rating system predicted winners more accurately than all other comparable computer-based systems[4] that were ranked in each year from 2013 to 2015. His ratings correctly predicted approximately 76 percent of games, which was better than the opening betting lines (although the midweek and updated betting lines were slightly more accurate). Keep in mind that Sagarin's ratings do not account for some information that oddsmakers use to set the betting lines, such as recent injury reports, player suspensions or weather.

His ratings were also used in the Bowl Championship Series (BCS), which determined the college football national championship game participants from 1998 to 2013. Sagarin said his involvement with the BCS was a blessing and a curse. The NCAA wanted to use his rating system to help pick the teams for the national championship but with a caveat.

"The NCAA told me, 'We know you need to use scores,'" Sagarin said. "'We're all coaches, we know the score tells you a lot. But as the NCAA, we can't officially have a rating system that uses the scores. Our official system can only take into ac-

count winning and losing.' They initially didn't even take into account home and away games!"

As a result, Sagarin came up with a system that ignored the actual scores of games, focusing almost entirely on wins and losses. He referred to it as the "Elo" system because of its similarity to the chess ratings developed by Arpad Elo in 1950.

Certainly at this point, actuaries from all disciplines can likely relate to Sagarin's dilemma. With almost any predictive modeling technique, using more information should generally yield more accurate results, if the modeling is done responsibly (for instance, by avoiding overfitting). However, there are often reasons certain variables need to be excluded. For instance, under President Obama's health care law, the Patient Protection and Affordable Care Act, health insurers on the individual exchange are only allowed to modify the base premium for an individual based on age, smoking status and geographic area.

Another commonality between Sagarin's current line of work and actuarial science is the blend of statistical competence and subject matter expertise. Sagarin has a variety of rating systems, but he began with a simple exponential smoothing system. Exponential smoothing is a technique for smoothing time series data. It uses all the historic information available but makes recent observations worth more than older ones. The actual balance of credibility between the recent observations and the older ones needs to be tuned for each forecast. It is possible to tune this balance by formulating a data-generating process and then minimizing errors, but just as often it is tuned by subjective domain knowledge and common sense.

In the case of Sagarin's rating system, the technique was used to modify the rating of a team on the basis of recent results, but it still required the smoothing factor to be chosen by the modeler to determine how much value to place on the recent results. Setting this factor appropriately is where Sagarin's intuition and knowledge of the game came into play.

"Let's say two teams play and you had them rated equally and one wins by two touchdowns. How much do you change it?" Sagarin said. "One person who knows nothing about sports would say, how about you put it to seven? Well, only a moron would do that. Teams don't change like that. You don't want to change the ratings like that. You only want to move it by a couple points."

This concept is not entirely different from the concept of actuarial credibility. If a health insurer offers coverage to a new large group, the premium is often set based on a "manual rate" based on the available information about the group, such as the demographic mix. However, at the start of the next year, the insurer likely would charge a premium that reflected a blend of the manual rate and the group's actual observed claims costs. The amount of weight given to the group's actual experience is referred to as the "credibility" of the group. Although there are commonly used formulas to estimate the credibility based on the size of the group and other factors, actuaries also have to rely on their own judgment.

Starting with an appropriate manual rate is critical to successful pricing in the insurance industry. A similar concept applies with sports ratings. Although there are some rating systems that are independent of the pre-season ratings, such as a Simple Rating System,[5] they can often produce unrealistic results early in the season. Other systems, such as those that rely on Bayesian concepts, rely heavily on the starting values (referred to as a priori estimates in a Bayesian framework).

"Bayesian systems will get better predictive results than going with pure unbiased results because [the pure unbiased results] are based only on the games early in the season," Sagarin said. "By mid-season, sort of 'All roads lead to Rome,' and all of the prediction systems are pretty similar."

Sagarin said the starting values are crucial to success. His starting values are based on a time-series analysis of each team's rating history.

"If you have good starting ratings, you'll have good ratings all year long," Sagarin said. "If you start off with Ohio State as the worst team in the country and Columbia as the best team, you're going to have problems."

Isaacs, a fellow of the Society of Actuaries who works with the IRS, said he "absolutely" sees similarities between Sagarin's work and the work we do as actuaries.

"After you've been doing this a while, you just have a sense that something's right or wrong," Isaacs said. "You just know. It's the same thing with his rating systems. You just have a sixth sense about what makes sense and what doesn't. I do that all the time. Someone will show me some numbers, and I'll say 'Something's not clicking, something just doesn't feel right.' I think there's a lot of carry-over to what Jeff does. He looks at the scores, but he looks at home and away, and all sorts of other things. To do what he does, he's got to be doing more than just plugging in scores."

I made a major career transition myself after working as a sportswriter in Columbus, Ohio, for more than three years. In 2009, I decided to begin taking actuarial exams and eventually was fortunate enough to secure a job at Milliman. At the time, I imagined the intersection of sports and actuarial science was virtually nonexistent. Seven years later, I can see I was wrong. The principles of actuarial science extend far beyond the insurance industry, as long as you know where to look. ■

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

ENDNOTES

1. To be precise, the point spreads are actually devised to entice an equal amount of money to be wagered on each team so the gambling organization is guaranteed to make a profit, after accounting for the fees they charge gamblers.
2. Ironically, given the same evidence, many gamblers would probably claim the Yankees couldn't possibly win eight consecutive Tuesday games against right-handed pitchers, and therefore they were "due" to lose.
3. *http://www.thepredictiontracker.com/predncaa.html.*
4. The term "comparable computer-based systems" excludes the actual betting lines and other systems that incorporate the betting line.
5. Doug Norris, "Simple Rating Systems: Entry-Level Sports Forecasting." Forecasting and Futurism (July 2015), *https://www.soa.org/library/newsletters/forecasting-futurism/2015/july/ffn-2015-iss11-norris-2.aspx.*

Seasonality of Mortality

By Kyle Nobbe

WARNING: Brace yourself for freshman year statistics. Modeling techniques such as neural networks, generalized linear models or random forests will not be found in this article.

Seasonality profoundly impacts mortality and that impact is widely felt throughout the life insurance industry. A variety of demographic, socio-economic and geographic factors influence the degree and direction of seasonal mortality. Tim Rozar explored these factors in great detail in 2012,[1] but the exceptional nature of excess mortality in the early part of 2015 begged an analysis to be conducted to further deepen our understanding.

Key outcomes from this research demonstrated:

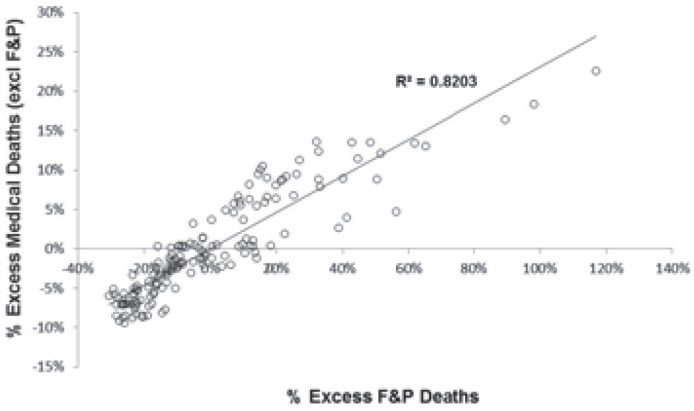
1. the strong correlation between influenza and pneumonia (F&P) related deaths to other causes of death, and
2. how F&P deaths are strong predictors of total population mortality using a simple linear regression.

F&P CORRELATION WITH OTHER CAUSES

Medical research demonstrates how individuals with pre-existing conditions[2] (such as diabetes, heart disease, etc.) are more likely to have serious complications when diagnosed with the flu or pneumonia. Knowing this fact, we wanted to test how excess F&P deaths in a given period correlated with other causes of death. To do this, ICD-10 codes were grouped into specific causes of death (such as F&P, unnatural, cancer, etc.) using the Centers for Disease Control and Prevention's Multiple Cause of Death Data. Figure 1 demonstrates how excess F&P deaths in a month correspond to an excess of all other medical deaths.

Clearly, a strong connection is evident between months with elevated F&P mortality and elevated mortality of other medical causes. Note that unnatural causes of deaths were excluded from the chart above due to their reverse seasonal nature of mortality. For a more granular look at which causes of death correlate the highest with excess F&P mortality, see Table 1 (below).

Figure 1: Scatterplot of F&P Excess Deaths by Month



Respiratory, cardiovascular, neurological, diabetes and other medical causes of death total about 70 percent of total U.S. population deaths, so a significant percentage of U.S. mortality is highly correlated to F&P mortality (which is about 2 to 3 percent of all U.S. deaths depending on the year).

PREDICTING TOTAL POPULATION MORTALITY

With the knowledge that F&P mortality correlates with the majority of other causes of deaths, we built a linear regression to predict total population mortality. We leveraged the CDC's National Center for Health Statistics (NCHS) data for this analysis

Figure 2: Prediction of Total Population Mortality

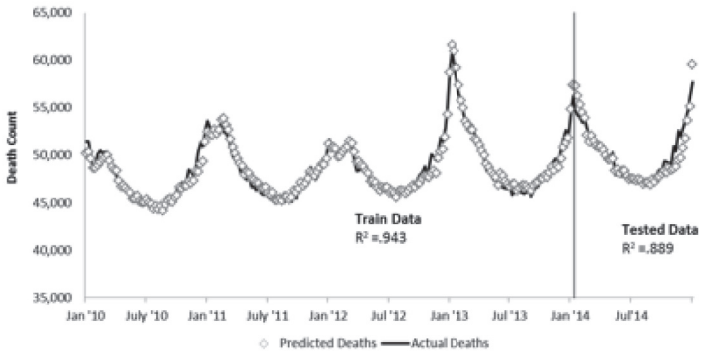


Table 1

CORRELATION BETWEEN EXCESS F&P MORTALITY AND OTHER CAUSES OF DEATHS							
	Respiratory (excl F&P)	Cardio	Other Medical	Neuro	Diabetes	Cancer	Unnatural
	0.95	0.90	0.87	0.87	0.87	0.37	-0.34

to incorporate more recent years of data. We fit the regression to weekly death counts between October 2009 and December 2013, which produced the following regression formula:

*Predicted Deaths = 31,648.07 + 4.077 * (Weeks since 10/1/2009) + 14.558 * (F&P Deaths)*

We tested the regression line on all 2014 deaths. Although the goodness of fit declined on the test data, the prediction was still strong. Figure 2 demonstrates the predictive power of using F&P deaths to predict to population mortality. I suspect the predictive power would only increase if cancer and unnatural deaths were stripped out.

IMPLICATIONS AND CHALLENGES

Monitoring flu season is widespread across numerous industries, and often the data and results are publically available. This includes hospitalizations, mortality, social media activity and search engine analytics. Insurance companies have an opportunity to better understand their business and stay ahead of potential epidemics thanks to the robustness of this data.

On the downside, there is still a lot of work to do. For starters, bridging the gap between population mortality and insured mortality can present challenges such as reporting lag, age standardization and underwriting wear-off, just to name a few. Additionally, flu and pneumonia forecasting is still in its infancy. Consider Columbia University's Prediction of Infectious Diseases model, which won the CDC's Predict the Influenza Season Challenge. The lead Columbia researcher commented, "Much work remains to improve the science of flu forecasting."

The life insurance industry has barely scratched the surface of this topic. I have no doubt advancements will continue to be made and it is imperative the industry be at the forefront of this important topic.

DATA SOURCES

- Centers for Disease Control and Prevention (CDC). Multiple Cause of Death 1999-2014 on the CDC Wide-ranging Online Data for Epidemiologic Research (WONDER) Online Database, released 2015. Data are compiled from data provided by the 57 vital statistics jurisdictions through the Vital Statistics Cooperative Program, *http://wonder.cdc.gov/mcd.html*.
- National Center for Health Statistics (NCHS). Weekly Mortality Surveillance Data, *http://www.cdc.gov/flu/weekly/nchs.htm*.

Kyle Nobbe, FSA, MAAA, is assistant actuary, global research and development, at RGA Reinsurance Company in Chesterfield, Mo. He can be reached at *knobbe@rgare.com*

REFERENCES

1. Rozar, Timothy. RGA. [Online] May 2012. *https://www.rgare.com/knowledgecenter/Documents/AW_SOA2012.pdf.*
2. CDC. Influenza - People at High Risk. [Online] *http://www.cdc.gov/flu/about/disease/high_risk.htm.*
3. Columbia University. Columbia Prediction of Infectious Diseases. [Online] *http://cpid.iri.columbia.edu/.*
4. CDC. Flu News & Spotlights. [Online] *http://www.cdc.gov/flu/news/predict-flu-challenge-winner.htm.*

Using Hadoop and Spark for Distributed Predictive Modeling

By Dihui Lai and Richard Xu

In the age of big data, the physical world we live in is dynamically mapped to the digital world in the form of data: news, messages, pictures, videos, health records, stock market data, you name it. Cloud computing and various sensors have made this process simpler than ever before. The ability to process enormous amounts of data in a timely and insightful manner is becoming the key to business success.

Computational power is essential in speeding up our data processing, and distributed computing systems (e.g., Hadoop, Spark) seem to be good candidates compared to many others (e.g., graphics processing units (GPUs), better central processing units (CPUs), quantum computers, etc.). On the other hand, predictive modeling (PM) has shown its importance in sophisticated data analysis (e.g., spam filters, product recommendations). A recent breakthrough in machine learning has also been the key to the success of Google's AlphaGo.[1]

However, the two components do not naturally proceed together. Modeling algorithms are focused on accuracy more than speed. Making them compatible with a distributed system requires a deep understanding of computer hardware, data structures and modeling mathematics. To an organization/company, this is simply translated into "cost." There may be less expensive ways to do it. In this article, we are going to review the ways to do scalable predictive analytics with an emphasis on open-source packages that support the Hadoop ecosystem.

IS YOUR JOB PARALLELIZABLE?

Perhaps one of the most important steps in moving a computing task to a distributed system is to determine if it can be parallelized and what the best parallelizing strategy could be. When building predictive models, there are mainly two computational intensive jobs: optimization and hyper-parameter search. Planning them well is critical to creating an efficient program.

In general, a machine learning algorithm or statistical model has an error function (sum of squared residuals, cross-entropy, etc.) it needs to minimize. The optimization algorithm updates the model parameters iteratively until the error function is minimized, considering some value (derivatives, predicting errors) estimated at each data point. A simple parallel strategy has two

steps: mapping each data point to the value-needs-estimate and adjusting the model parameters accordingly. This description might remind you of a map-reduce job and, indeed, the strategy can be easily implemented in Hadoop/Spark since map-reduce is well supported there.

Besides the optimization, statistical models normally have a list of hyper-parameters associated with them (e.g., distribution prior, sampling ratio, variable selection ratio, etc.) Determining the best hyper-parameters is critical to model accuracy, and searching through the hyper-parameter space is a common practice. The search process is computationally expensive, and speeding it up will allow searching a larger space. An intuitive solution here is to create a pool of models with promising hyper-parameters and distribute them to worker nodes for concurrent evaluation.

Writing parallel code is nontrivial. It is tricky to balance efficiency with the overhead the code will introduce. It is not uncommon for a developer to find that after days or weeks of diligent work, the map-reduce job he wrote helps little to none on a program's execution time. In the following sections, we will introduce some active open-source projects that aim to make scalable machine learning easy.

SCALABLE MACHINE LEARNING PACKAGES

MLLIB/SPARKNET

More people are now accepting Spark as the new process engine for the Hadoop ecosystem.[2] Spark's in-memory support has made it ideal for developing scalable machine learning algorithms. MLlib is a product of such efforts from the Spark

community. The library covers a wide range of common algorithms: linear regression, naïve Bayes, decision trees, k-mean, etc. (see Table 1). SparkNet, the all-star deep-learning algorithm, is not included in MLlib but was developed in a separate Spark package.[3]

The library conveniently provides APIs to languages like Python, Java and Scala. As the library is built on top of built-in data structures like RDD or data frames, Spark's data processing tools (e.g., Spark SQL) come in handy to the user. Data manipulations like merging or subsetting can be handled smoothly without much painstaking work.

However, there is one piece missing in the MLlib that is important for actuarial use—the generalized linear model (GLM). Although linear regression and logistical regression are supported, MLlib is missing two important members of the GLM family: the Poisson distribution and the Tweedie distribution. These distributions are responsible for frequency models and loss-cost models.

Table 1: Comparisons of the machine learning algorithms supported by H2O, MLlib/SparkNet and Mahout

	H2O	MLlib/SparkNet	Mahout
Generalized Linear Model	X		
Random Forest	X	X	X
Naïve Bayes	X	X	X
Gradient Boosting Machine	X	X	
K-Mean Clustering	X	X	X
Cox Proportional Hazards	X		
SVM		X	

H2O

Compared to MLlib, which might seem like a direct application of the Spark engine, H2O was aiming to solve scalable statistical problems with its creation. As a key to fast machine learning algorithms, H2O supports in-memory processing as well. To actuaries' delight, H2O does support GLM and includes distributions like Poisson, gamma and Tweedie. Moreover, H2O also supports survival analysis like Cox-model (Table 1). However, H2O is slightly weak in data manipulation. For example, to add a derived variable from an existing column, users have to write a map-reduce job for the H2O-frame.

H2O can be plugged into Hadoop or Spark (with sparkling-water) clusters easily and leverages the capabilities of the distributed system: resource management, HDFS storage, data manipulation, etc. The current version of sparkling-water supports Scala and Python. R users can install H2O as a library and use H2O cluster by connecting to the service.

MAHOUT

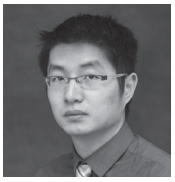
Apache Mahout has a slightly longer history than the two packages described above. Most of its algorithms are designed within the framework of map-reduce. The initial project has been focused on algorithms like clustering and classification. In light of the Spark success, the Mahout project has recently shifted its focus from writing map-reduce algorithms to providing a platform supporting H2O, Spark and Apache Flink.

OTHER

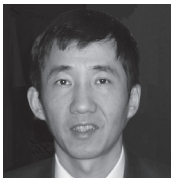
Besides the open source projects listed above, commercial software like SAS, Revolution R (Microsoft) and Big R (IBM) all provide scalable predictive modeling on Hadoop/Spark with nontrivial cost—as the size of the cluster goes up, the cost will increase proportionally.

DISCUSSION

As the era of big data approaches, the need for fast big data analytics is becoming greater than ever. The open source projects we reviewed here provide us ways to gain power at relatively low cost. However, the packages are created with their own flavors and each has features others do not. Depending on the application, users need to choose the one that best fits their need. If your PM application requires lots of data manipulation, MLlib could be the best option. If the application requires using a model like GLM, H2O is your best friend. And, if your organization has plenty in its budget, it is hard to say no to the commercial software! ■



Dihui Lai, Ph.D., is assistant data scientist, global research and development, at RGA Reinsurance Co. in Chesterfield, Mo. He can be reached at *dlai@rgare.com.*



Richard Xu is vice president and actuary, head of data science, at RGA Reinsurance Co. in Chesterfield, Mo. He can be reached at *rxu@rgare.com.*

ENDNOTES

1. David Silver, et al., "Mastering the Game of Go With Deep Neural Networks and Tree Search," Nature 529, no. 7587 (2016), 484-89.
2. Phillipp Moritz, Robert Nishihara, Ion Stoica and Michael I. Jordan, "SparkNet: Training Deep Networks in Spark," (conference paper, ICLR, 2016).
3. Tom White, Hadoop: The Definitive Guide, 3rd ed. (Sebastaopol, CA: O'Reilly Media/Yahoo Press, 2012).

Regression and Classification: A Deeper Look

By Jeff Heaton

Classification and regression are the two most common forms of models fitted with supervised training. When the model must choose which of several discrete classes the sample data belong to, classification is used. Similarly, when the model must compute a numeric output from the input data, regression is used. Classification is used when the output is discrete, or categorical, and regression is used when the output is continuous, or numeric.

It quickly becomes more complex than this simple case. Many models, such as support vector machines or generalized linear models (GLMs), only support binary classification—they can only directly classify between two discrete classes. Yet these models are often used for many more than two classes. Similarly, neural networks and linear regression only directly support regression. This article will look at three distinct applications of supervised learning:

- binary classification
- multi classification
- regression

The exact means by which several models support these three will be discussed. This article will specifically examine the following models:

- generalized linear regression (GLM)
- linear regression
- neural networks
- support vector machines
- tree-based models

SIMPLE REGRESSION

Linear regression is one of the most basic, yet still useful, types of model. One representation of the linear regression formula is given by Equation 1.

Equation 1: Linear Regression

$$\hat{y} = \beta_1 x_1 + \cdots + \beta_n x_n$$

The output prediction $\hat{y}$ (y-hat) is calculated by the summation of multiplying each input element (x) by a corresponding coefficient/weight (β). Training the linear regression model is simply a matter of finding the coefficient values that minimize the difference between $\hat{y}$ and the actual y. It is very common to append a constant value, typically 1, to the input vector (x). This constant allows one of the coefficient (β) values to serve the y-intercept.

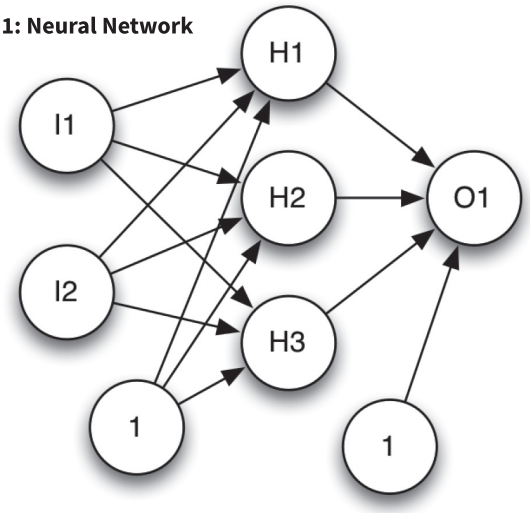
The returned value is numeric—a regression was performed. Common examples of linear regressions derive coefficients to determine shoe size, based on height, or a person's income based on several other numeric observations. Linear regression is best at modeling linear relationships. For nonlinear relationships, a neural network or generalized linear model (GLM) might be used. A single neuron in a neural network is calculated by Equation 2.

Equation 2: GLM or Single Neuron Calculation

$$\hat{y} = f(x, w) = \Phi(x_1 w_1 + \cdots + x_n w_n)$$

The output from a single neuron is very similar to the linear regression. An input/feature vector (x) is still the primary input. However, neural network terminology usually refers to the coefficients (β) as weights (w). Usually a constant input term is appended, just like linear regression. However, neural networks terminology refers to this weight as a bias or threshold, rather than the y-intercept. The entire summation is passed to a transfer, or activation function, denoted by Φ. The transfer function is typically sigmoidal, either logistic or the hyperbolic tangent. Newer neural networks, particularly deep neural networks, will often use a rectifier linear unit (ReLU) as the transfer function. Neural networks are typically made of layers of neurons, such as Figure 1.

Figure 1: Neural Network



The output from a neural network is calculated by first applying the input vector (*x*) to the input neurons, in this case I1 and I2. A neural network must always have the same number of input neurons as the vector size of its training data (*x*). Next, calculate the values of each hidden neuron H1, H2, etc., working forward until the output neuron(s) are calculated.

The output for a GLM is calculated exactly the same as a single neuron for a neural network. However, the transfer/activation function is referred to as a link function. Because of this, a neural network can be thought of as layers of many GLMs.

The error for a neural network or GLM can be thought of as the difference between the predicted output ($\hat{y}$) and the expected output (y). A common measure of the error of neural networks, and sometimes GLMs, is root mean square error (RMSE), the calculation for which is shown by Equation 3.

Equation 3: RMSE

$$E = \sqrt{\frac{\sum_{i=1}^{N}(\hat{y_t} - y_t)^2}{N}}$$

The constant N represents the number of items in the training set. RMSE is very similar to the standard deviation calculation used in statistics. RMSE measures the standard deviation from the expected values.

BINARY CLASSIFICATION

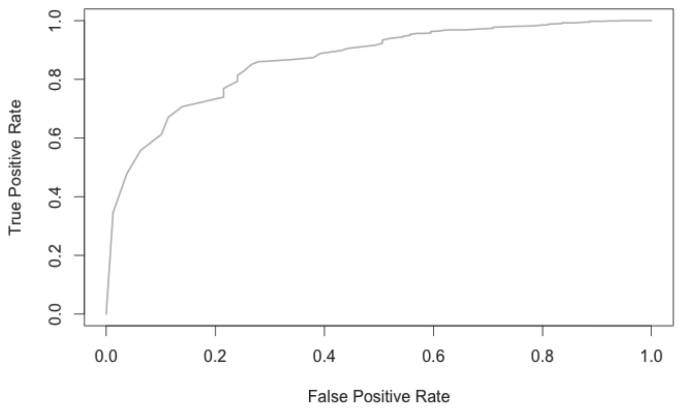
Binary classification is when a model must classify the input into one of two classes. The distinction between regression and binary classification can be fuzzy. When a model must perform a binary classification, the model output is a number that indicates the probability of one class over the other. This classification is essentially a regression on the probability of one class vs. the other being the correct outcome! For many models, binary classification is simply a special case of regression.

A popular form of binary classification for the GLM model is logistic regression, where the link function is the logistic function. If the GLM, using logistic regression, is to classify more than two categories, a special voting arrangement must be used. This is discussed later in this article.

Binary classification provides a number that states the probability of an item being a member of a category. However, this brings up the question of what a sufficient probability is for classification. Is a 90 percent probability enough? Perhaps a 75 percent probability will do. This membership threshold must be set with regard to the willingness to accept false positives and false negatives. A higher threshold decreases the likelihood of a false positive, at the expense of more false negatives. A receiver

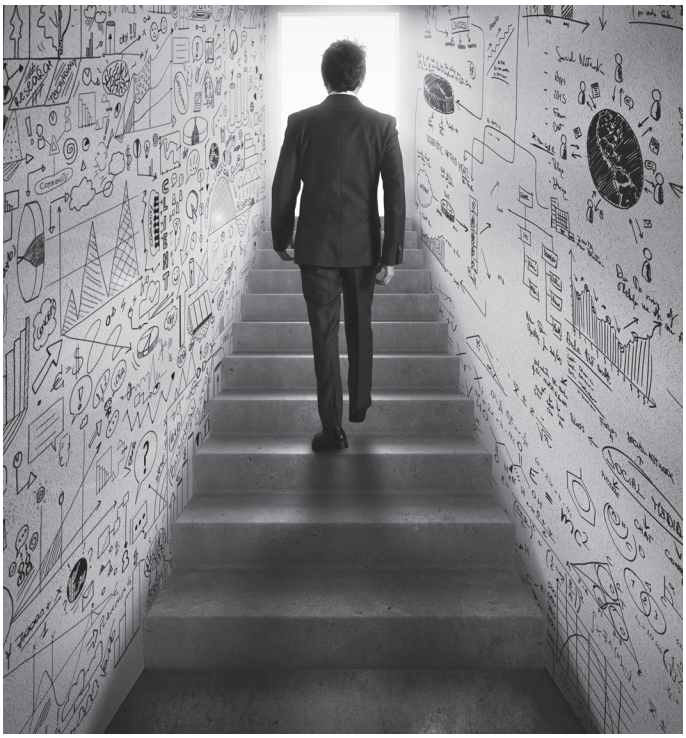
operating characteristic (ROC) curve is often used, for binary classification, to visualize the effects of setting this threshold. Figure 2 shows a ROC curve.

Figure 2: ROC Curve



As the threshold is set more or less restrictive, the true positive rate and false positive rates change. A threshold is represented as one point on the curved line. If the true positive rate is very high, so will be the false positive rate. The reverse also holds true.

Sometimes it is valuable to measure the effectiveness of the model, independent of the choice of threshold. For such cases, the area under the curve (AUC) is often used. The larger the



area below the curve, the better the model. An AUC of 1.0 is a perfect, but highly suspicious, model. Nearly "perfect" models are rare, and usually indicate overfitting. Calculating the AUC can be complex and often employs similar techniques to integral estimation. It is rare that AUC calculation can be performed using definite symbolic integration.

MULTIPLE CLASSIFICATION

AUC curves are only used for binary classification. If there are more than two categories, a confusion matrix might be used. The confusion matrix allows the analyst to quickly see which categories are often mistaken for each other. A confusion matrix for the classic iris dataset is shown by Figure 3.

Figure 3: Iris Confusion Matrix

		Predicted		
		Setosa	Versicolor	Virginica
Actual	Setosa	46	1	3
	Versicolor	2	46	1
	Virginica	1	1	48

The iris dataset is a collection of 150 iris flowers, with four measurements from each. Additionally, each iris is classified as one of three species. This dataset is often used for example classification problems. The confusion matrix shows that the model in question predicted Setosa correctly 46 times, but misclassified a Setosa as Versicolor once, and Virginica three times. A strong model will have its highest values down the northwest diagonal of a confusion matrix.

It is important to understand how a model reports the prediction for a multiclassification. A model will report a vector for the input data. The model might report 90 percent Setosa, 7 percent Versicolor, and 3 percent Virginica for a set of flower measurements that the model felt was likely Setosa. In this case, the model would return the following vector:

[0.9,0.07,0.03]

This is very different from the typical multiple choice question format that might be seen on an actuarial exam. For such an exam, the answer must be chosen as either A, B, C or D. The model has the advantage of being able to choose its confidence in each of the possible choices.

Classification problems are often numerically evaluated using the multiple log loss, as shown by Equation 4.

Equation 4: Multi-Log Loss

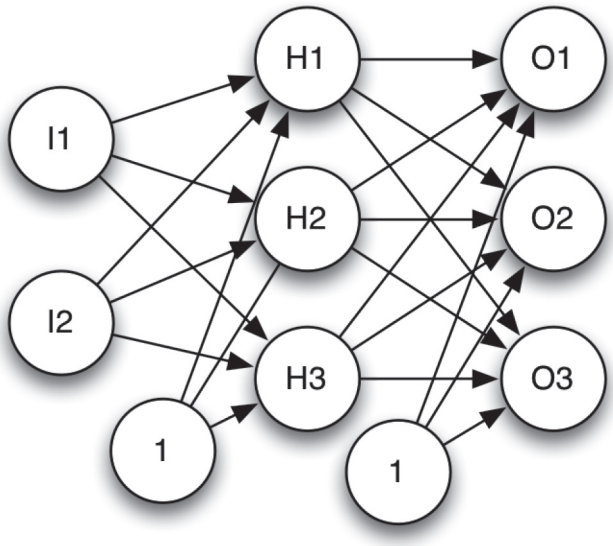
$$E = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{i,j}\log(\hat{y}_{i,j})$$

The constant N represents the number of training set items and M represents the number of classes. Like previous equations in this article, y represents the model prediction and ŷ represents the expected outcome. The lower the log loss, the better. To explain how Equation 4 works, think of the multiple choice exam previously mentioned. If the correct answer for a particular question was A, and the model had given a .97 probability to A, then -log(0.97) points would be added to the average score. Log loss can be harsh. Predicting 1.0 correctly will add zero log-points to the error, but predicting 1.0 incorrectly will give an infinitely bad score. Because of this, most models will never predict 1.0.

MULTIPLE REGRESSION

Just like multiple classification, multiple regression also exists. Neural networks with multiple outputs are multiple regression models. Usually, a neural network with multiple outputs is used to model multiple classification. This is how neural networks, which are inherently regressive, are made to support classification. A binary classification neural network simply uses a single output neuron to indicate the probability of the input being classified into one of the two target categories. For three or more categories, the output neurons simply indicate the class that has the greatest probability. Figure 4 shows a multiple output neural network.

Figure 4: Multi-Output Neural Network



Because a binary classification neural network contains a single output neuron, and a three or more classification network would contain a count equal to the number of classes, a two-output neuron neural network is rarely used. While a neural network could be trained to perform multiple regressions simultaneously, this practice is not recommended. To regress multiple values, simply fit multiple models.

SOFTMAX CLASSIFICATION

For classification models, it is desired that the probabilities of each class sum to 1.0. Neural networks have no concept of probability. The output neuron, with the highest value, is the predicted class. It is useful to balance the output neurons of neural networks, and some other models, to mirror probability. This is accomplished with the softmax, as shown by Equation 5.

Equation 5: Softmax

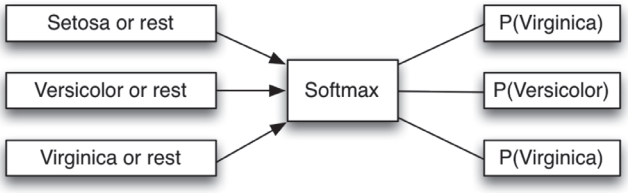
$$\sigma(x)_j = \frac{e^{x_j}}{\sum_{k=1}^{K} e^{x_k}} \text{ , } j = 1, \ldots, K.$$

Softmax can be used to transform any multi-output regression model to a classification model. Most neural network-based classifications make use of the softmax function. It is based on the logistic function and provides the same sort of squashing effect at the extremities. The softmax function is very similar to simply summing the output neurons and balancing each neuron to become the proportion of this summation. This approach is often called normalization or simply hardmax. The softmax softens this approach and usually makes the probabilities more realistic.

VOTING

Many models can only function as binary classifiers. Two such examples are GLMs and support vector machines (SVM). Not all models have this limitation; any tree-based model can easily classify beyond two classes. For a tree, the leaf-node specifies the class. It is very easy to convert any binary classifier into a three or more classifier. Figure 5 shows how multiple binary classifiers could be adapted to the iris dataset for classification.

Figure 5: Model Voting



Essentially, a classifier is trained for each of the output categories. For the iris dataset, three additional datasets are created, each as a binary classifier dataset. The first dataset would predict between Setosa and all other classes. Each class would have a dataset and model that predicts the binary classes of that category and all others. When using such a model, the input data would be presented to each of the three models, and the data would be classified as belonging to the class that predicted the highest probability.

Like a multi-output neural network, it would be helpful if the probabilities of each class summed to 1.0. This can be accomplished with the softmax function. By using multiple binary classifiers and a softmax, any binary classifier can be expanded beyond two classifications.

CONCLUSIONS

Classification and regression are the two most common formats for supervised learning. As this article demonstrated, models have entirely different approaches to implementing classification and regression. Often the software package will take care of these differences. However, understanding the underpinnings of the models can be useful. For example, if a model were trained to recognize a large number of classes, then a GLM or SVM might not be a good choice. If there were 10,000 possible outcome classes, a binary-only classifier would need to create a voting structure of 10,000 models to vote upon each classification. A large tree/forest or neural network might be able to more effectively handle such a problem. ■

Jeff Heaton is the author of the "Artificial Intelligence for Humans" series of books, and senior data scientist at RGA Reinsurance Co. in Chesterfield, Mo. He can be reached at *jheaton@rgare.com.*

From Deep Blue to DeepMind: What AlphaGo Tells Us

By Haofeng Yu

Early February in 2016, Demis Hassabis, one of Google DeepMind's founders, tweeted: "Thrilled to officially announce the 5-game challenge match between #AlphaGo and Lee Sedol in Seoul from March 9th-15th for a $1M prize!" While Hassabis was a name I barely knew and AlphaGo sounds like another of Google's toys with a catchy name, growing up playing Go, I knew about Lee very well. The Korean professional Go player had been at the top of the game for almost a decade. The 18 world championships he collected are nothing short of Roger Federer's 17 or Tiger Woods' 14 grand slam titles in their respective fields, tennis and golf. The competition didn't seem to be a good match-up. "I would bet anything that AlphaGo won't go anywhere," I told my friends.

The competition took place in Seoul as scheduled. To the surprise of many fans, including Go professionals, AlphaGo beat Lee four games to one, with the human's sole win coming from the fourth game, merely a consolation that doesn't matter in the best-of-five setting. This is stunning, devastating, yet equally interesting. It inevitably reminds people of the chess match that took place in 1997 between IBM's super computer Deep Blue and Garry Kasparov, the reigning champion at that time. Deep Blue won.

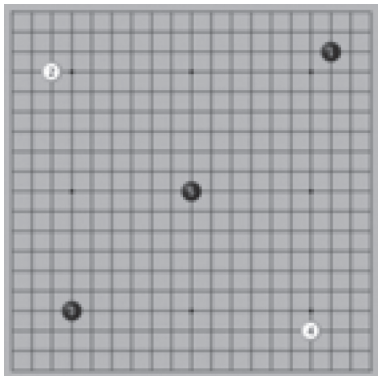
Humanity's intellectual pride continued to be humbled with IBM's Watson beating two champs on the game show "Jeopardy!" in 2011, and now AlphaGo winning at Go, the game many applaud as the final line of defense of human intelligence. Many questions ensue, including: What are DeepMind and AlphaGo? What can AlphaGo tell us, particularly, actuaries? To begin with, let's talk about Go.

ABOUT GO

Have you seen the 2001 movie "A Beautiful Mind"? There is a scene at the beginning where John Nash (Russell Crowe) awkwardly wanders around Princeton's campus, "extracting an algorithm to define [the] movement" of pigeons, while making notes. Very soon, he is dragged into a game of Go, with one semester's free laundry service at stake. Nash loses and claims the game is flawed and his perfect move was ruined.[1] Had he extracted some algorithms for Go instead of for the birds, like AlphaGo did, he could have enjoyed the free laundry service.

The game of Go originated in China more than 2,500 years ago. The rules are simple: Players take turns placing black or white stones on the board, a 19-by-19 square grid, trying to capture the opponent's stones or surround empty space to mark as their own territory. See Figure 1.

Figure 1: The Game of the 20th Century: Go Seigen (Black) vs. Honinbo (White) 1933



Source: Unknown

As simple as the rules are, Go is a game of profound complexity. Its abstract concept of "shi," sensible but indescribable, is unique to the game and often linked to Oriental philosophy, even at national strategic level.[2] Unlike Western chess, which has about 40 moves in a game, Go can last for up to 200.

According to Google DeepMind's site, "There are more possible positions in Go than there are atoms in the observable universe. … Go is played primarily through intuition and feel, and because of its beauty, subtlety and intellectual depth, it has captured the human imagination for centuries."[3]

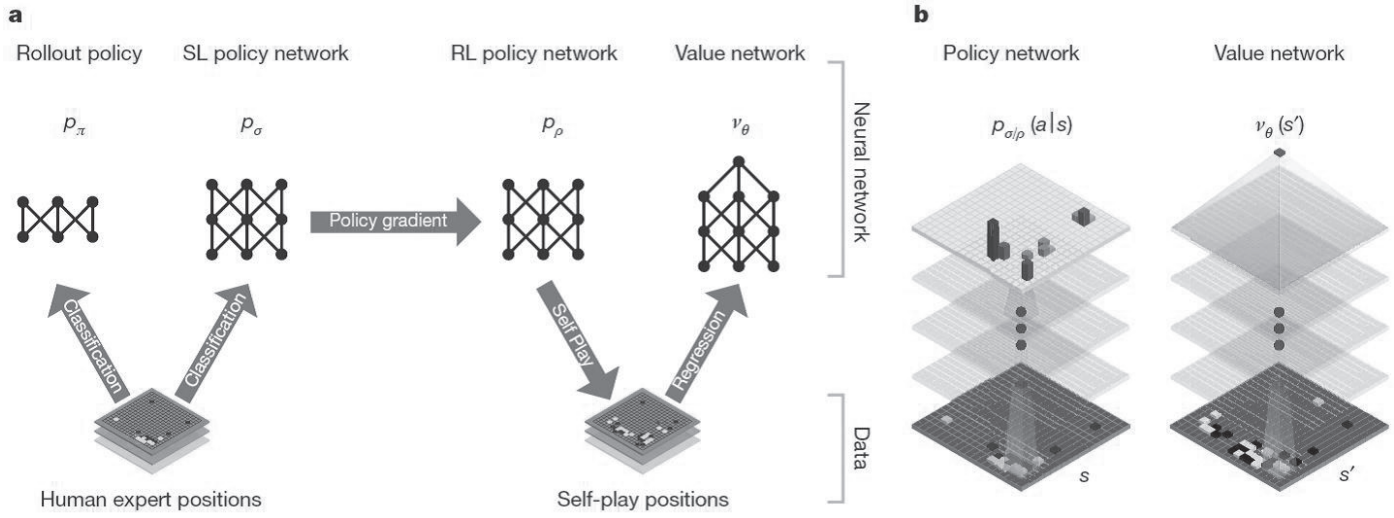
Two quotes from 20th century Chess and Go player Edward Lasker summarize chess and Go this way:

"It has been said that man is distinguished from animal in that he buys more books than he can read. I should like to suggest that the inclusion of a few chess books would help to make the distinction unmistakable." – *The Adventure of Chess*

"While the Baroque rules of Chess could only have been created by humans, the rules of Go are so elegant, organic, and rigorously logical that if intelligent life forms exist elsewhere in the universe, they almost certainly play Go."

No wonder there are so many efforts, including from Facebook, to build a Go application—it simply offers higher levels of—if not the ultimate—challenge. It was thought it would be at least

Figure 2: Neural network training pipeline and architecture



Source: Silver et al., "Mastering the Game of Go."

another 10 years before a machine could beat a human professional in Go; it happened much more quickly.

ABOUT DEEP BLUE

Back in 1997, how did Deep Blue beat Kasparov, the reigning world champion? IBM explains on its Deep Blue website, "The answer lies in its unique combination of innovative software engineering and massive parallel processing power."[4]

To the first point, the keys to IBM's software engineering are tree search with alpha-beta pruning technique and hand-crafted evaluation functions, which do not necessarily represent advanced mathematics or a heavy use of statistics!

To the second, Deep Blue is a massively parallel "32-node IBM RS/6000 SP high-performance computer … capable of evaluating 200 million positions per second"; now we know this kind of computing power can be available in each household.

While Deep Blue attains its strength more or less out of brute force computing power, back in the day, it was a modern marvel.

ABOUT DEEPMIND AND ALPHAGO

Founded in Britain in 2010, the artificial intelligence company Google DeepMind was acquired and renamed by Google in 2014. Google describes AlphaGo as a computer Go that combines Monte Carlo tree search with deep neural networks that have been trained by supervised learning (SL), from human expert games, and by reinforcement learning (RL) from games of self-play.[5] The AlphaGo team also published a paper in Nature in January 2016, which offers comprehensive technical details, for your academic curiosity.[6]

As mentioned earlier, since the search space of future moves of a Go game is so large that no AI can explore every possibility, how did AlphaGo accomplish the mission impossible? Figure 2 tells you where AlphaGo derives its amazing playing strength.

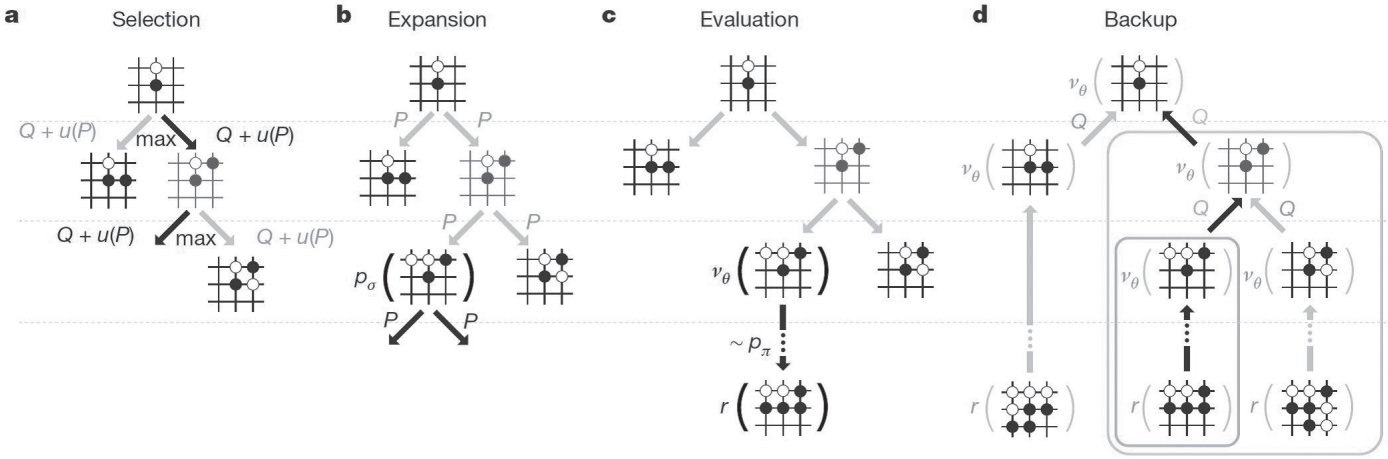
To the best of my understanding, its secret power comes from the following four elements.

1) **Policy networks** (AlphaGo's left brain). Given the current situation, these networks predict moves that human experts would likely pick. There are two kinds, or phases:

   - Supervised learning. The policy network was trained with numerous information, 30 million positions, from Go games that had been played by human experts; it predicts by maximizing the likelihood of human expert moves.
   - Reinforcement learning. The policy network was training by playing "against itself" millions of times, in a sense teaching itself which moves and strategies worked and which didn't; it predicts by maximizing expected outcomes (of winning).

2) **Rollout policy** (AlphaGo's legs and hands, as it "acts" without thinking/using its "brains"). Given the current situation, this policy predicts moves that human experts would make, similarly to policy networks, but with much

Figure 3: Monte Carlo tree search in AlphaGo



Source: Silver et al., "Mastering the Game of Go."

greater speed. It plays in a way akin to "intuition," so as to achieve a balance between accuracy and speed.

3) **Value network** (AlphaGo's right brain). Given the current situation, it evaluates and spits out the odds of winning or losing. With this function, AlphaGo is able to evaluate its moves quantitatively. Generally speaking, the value function of Go is highly nonsmooth and irregular.

4) **Monte Carlo tree search** (AlphaGo's body). This is the framework that integrates all the parts.

In a training pipeline, the AlphaGo team "pass in the board position as a 19 × 19 image and use convolutional layers to construct a representation" and then "use neural networks to reduce the effective depth and breadth of the (Monte Carlo) search tree (4), evaluating positions using a value network (3), … sampling actions using a policy network (1)," and balancing speed and accuracy with the fast rollout policy (2).

None of the four pieces is utterly new; however, the integration of these concepts, in such a creative and efficient way, is a work of beauty.[7]

WHAT DOES ALPHAGO TELLS US?

The first lesson is that while computing power is still important, its weight has declined. Back in 1997, IBM touted its computing power as one major contributing factor; in 2016, DeepMind seems to intentionally refrain from using super power. The AlphaGo that defeated Lee was a distributed version that uses 1,202 central processing units (CPUs) and 176 graphics processing units (GPUs). Given Google's capacities, it can certainly come up with a stronger AlphaGo if they wish.

The second lesson is that the brute force of Deep Blue has evolved into a whole new form, in the name of machine learning. Unlike Deep Blue, which employed exhaustive tree search with alpha-beta pruning, AlphaGo learns things "brute-forcely" from scratch. In a sense, the brute force is manifested by its "diligence"—AlphaGo mimics an extremely diligent, but not necessarily genius, student who is willing to learn from millions of human's play and self-play, tediously.

The third lesson we take away here is that data is the key. Deep Blue relied on a huge database of hand-crafted books on openings and endgames to simplify its search; without the daunting 30 million human positions AlphaGo has learned, I doubt the reinforcement learning by self-play can add much value and AlphaGo's strength shall be discounted.

I believe these points, especially the third one, are particularly important for us actuaries. While we have started seeing so-called "disruptive innovations" of machine learning and predictive analytics in our work, without high quality and business specific data, anything that they mean could be misleading. So, for companies who strive to automate their agency, underwriting and claims, or even investment and asset liability management (ALM) processes, they had better invest in data, so as to save for a rainy day.

WHAT HAS ALPHAGO NOT TOLD US YET?

First, the new form of brute force mentioned above may be easily translated into other logic-based territories, not limited to games. True, AlphaGo can only play Go right now. It cannot even move one stone by itself—one of its creators, Ajay Huang, had to sit in front of Lee Sedol and place stones on its behalf.

But its way of learning, guided by minimal hand-crafted rules, is truly inspiring.

Second, there exists another powerful and relevant machine learning tool that has not been mentioned yet—unsupervised learning (UL). Judging from the paper in Nature, AlphaGo doesn't seem to have been trained by UL, or at least, DeepMind didn't make it explicit. But some of AlphaGo's moves are far from we humans' play book. For example, the 37th move in Game 2—no human player would play like this; yet, it was a key play whose importance only was revealed after 20 more exchanges. Its own way of playing! One has to wonder, if DeepMind does train AlphaGo using UL, can it teach humans even more?

Interestingly enough, but I bet that DeepMind won't be satisfied by producing merely top video game or Go players. We have reason to believe that DeepMind and its competitors are aiming for more, especially in this era when big data, machine learning, cloud computing, Internet of things (IoT), augmented reality (AR) and virtual reality (VR) bring our physical world closer than ever to virtual worlds. While AlphaGo-like "narrow" AIs (as described by Hassabis) are still far away from their ultimate form, artificial general intelligence (AGI), they are marching in that direction.

NOT JUST FOR APRIL FOOLS'

In Davos-Klosters, Switzerland, this January, the fourth industrial revolution, or Industry 4.0, driven by rising usage of big data and artificial intelligence in all aspects of the economy, emerged as one of the main themes at the 46th World Economic Forum. One report predicts "7.1 million redundancies by 2021, mainly in the fields of management and administration, particularly in the healthcare sector." About the same time, McKinsey released its outlook that automated systems may take over up to 25 percent of insurance jobs.[8]

On the other end, DeepMind just announced on their website that it had struck a deal collaborating with the U.K.'s National Health Service; IBM revealed on their website plans to move into telehealth and telecare five years after IBM Watson toppled the game show "Jeopardy!" Granted, the health and insurance industry is not the only space where actuaries live, but it has been our natural habitat!

Coincidentally, or intentionally in light of the AlphaGo hype, a friend shared with me a news item with the headline "First Robot Run Insurance Agency Opens for Business"—what a "classic" teaser by a "classic" name, Lirpa Loof, on an April Fools' Day! Somehow, it appears not just for April Fools'.

STILL A LONG WAY TO GO

Not all AIs succeeded in challenging humans. Claudico, an AI from Carnegie Mellon University (CMU), lost the Brains vs. Artificial Intelligence challenge in a type of Texas hold 'em poker game in 2015. Interestingly, CMU is also the birthplace of Deep Blue.

In summary, here are two takeaway messages for AI.

- Be humble; even a sophisticated game like Go may represent only a limited and partial perspective of the human unpredictable nature.
- Get more training, supervised or unsupervised, on bluffing.

"Right now, humans are doing OK,"[10] said Doug Polk, a former World Series of Poker champion, who just "defeated" Claudico.

The author would like to thank Aolin Zhang for sharing the April Fools' news and helpful discussion. ■

Haofeng Yu, FSA, Ph.D., is actuary and director, of Inforce Management at AIG. He also serves as webcast and research coordinator of the Predictive Analytics and Futurism Section Council. He can be reached by *Haofeng.Yu@aig.com.*

REFERENCES

1. "The Challenge," Beautiful Mind, directed by Ron Howard, 2001, YouTube clip, 2:26, posted by "andStack," Oct. 8, 2010, *https://www.youtube.com/watch?v=GmlSSSN7C78&nohtml5=False.*
2. David Lai, "Learning from the Stones: A Go Approach to Mastering China's Strategic Concept, Shi" (Advancing Strategic Thought Series, Strategic Studies Institute, May 2004), *http://www.strategicstudiesinstitute.army.mil/pubs/display.cfm?pubID=378.*
3. "The Game of Go," Google DeepMind, accessed May 16. 2016, https://www.deepmind.com/alpha-go.
4. "How Deep Blue Works: Under the Hood of IBM's Chess-Playing Supercomputer," IBM, accessed May 16, 2016, *https://www.research.ibm.com/deepblue/meet/html/d.3.2.html.*
5. "AlphaGo: Using Machine Learning to Master the Ancient Game of Go," Google (blog), Jan. 27, 2016, *https://googleblog.blogspot.nl/2016/01/alphago-machine-learning-game-go.html.*
6. David Silver, et al., "Mastering the Game of Go With Deep Neural Networks and Tree Search," Nature 529 (Jan. 28, 2016), *http://www.nature.com/nature/journal/v529/n7587/full/nature16961.html.*
7. While the fully armed AlphaGo deserves all compliments, let us not underappreciate the fineness of each part—they are refined by DeepMind to a level such that each standalone piece can challenge other competitor Go programs in the world.
8. Unknown,"Robots put five million jobs at risk," SwissInfo, accessed May 24, 2016, *http://www.swissinfo.ch/eng/world-economic-forum_digital-revolution-puts-five-million-jobs-at-risk/41900634.*
9. "How Automation Will Whack Up to 25% of Insurance Jobs: McKinsey", Insurance Journal, accessed May 24, 2016, *http://www.insurancejournal.com/news/national/2016/02/01/397026.htm.*
10. Noah Bierman, "Artificial Intelligence Bot vs The Poker Pros," LATimes, accessed May 24, 2016, http://www.latimes.com/nation/great-reads/la-na-c1-claudico-poker-20150521-story.html.

Exploring the SOA Table Database

By Brian Holland

The Society of Actuaries database has historical values of several types of tables. This article goes through some basic data exploration techniques to show how different approaches look. Here I'm aiming for a quick view into the tables that are vectors, such as lapse rates by duration or ultimate mortality rates. We could also deal with matrices such as select and ultimate tables by laying rows out end to end to make a longer vector.

When would you do this type of thing in practice? You might if you had thousands of tables installed into a valuation system or pricing repository and you wanted to look for features. Those features could conceivably include typos, which should stand out and be caught.

There were 3,909 vectors among the 2,621 table files extracted from the SOA database. Some table files included two or more vectors. The winner was No. 1531,[1] which has 55 vectors of durational lapse rates by different segments of business. Of course, the rates could be organized differently than as a loose collection of vectors. However, the purpose here is to skip all organizational points and look quickly at the data as they are expressed in the database. Missing values are plugged with zero for that purpose, and different axes are lined up: durations in some cases, or ages in others. There are 141 dimensions: the longest vector has 127 values, but some only overlap, and the vectors go from 0 (like some attained ages) to 140 (a Brazilian mortality table[2]).

DIMENSION REDUCTION: WHAT IT IS

We are all intuitively familiar with some dimension reduction. Shadows reduce a 3-D object to 2-D; if the shadow is on a stick, the dimension drops from 3-D to 1-D. I find it helpful to imagine dimension reduction as rotation of a higher-dimensional object in a way to cast the widest shadow. The object does not have to just be three dimensions; here we reduce 141-dimensional objects to two dimensions. We will miss many facets of the data, but it is a start to get a view.

Singular value decomposition (SVD) is the dimension reduction technique used here. If we imagine each vector of the 3,909 in 141 dimensions, what we're doing with SVD is rotating the 141 axes so the first two rotated axes catch the biggest shadow, or dispersion of the points. There are many proper mathematical treatments on the web, which are good excuses to bone up on linear algebra. Note that principal component analysis (PCA) is closely related.

In Figure 1, each dot represents one of the 3,909 vectors. Around (0,0) we have most of the points; there are just a few outliers.

Figure 1: Plot of tables by first two right singular vectors (main dimensions)

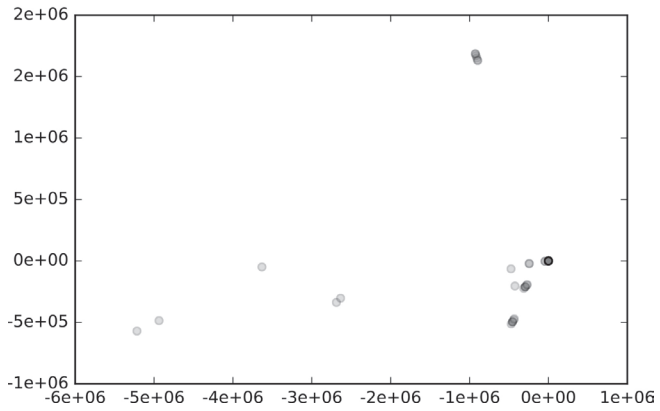
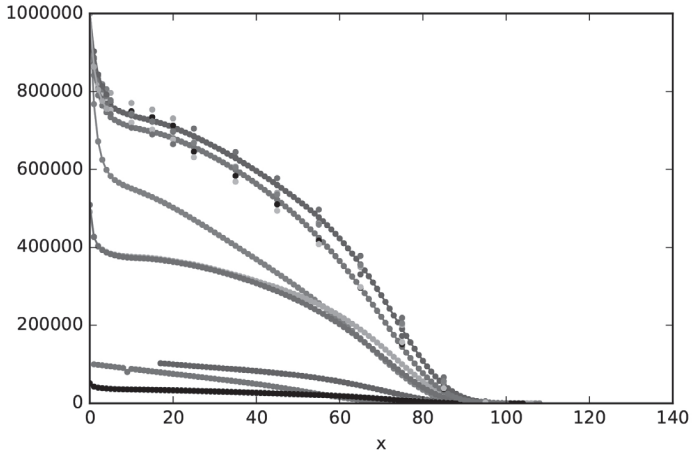


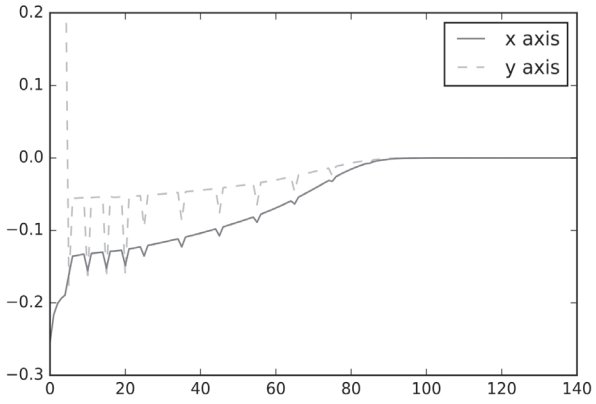
Figure 2 shows the vectors that the outlying points in Figure 1 represent. Those vectors are mostly English and Scottish life tables. It's no wonder they're outliers in Figure 1; they look nothing like mortality tables.

Figure 2: Actual main outlying vectors form Figure 1



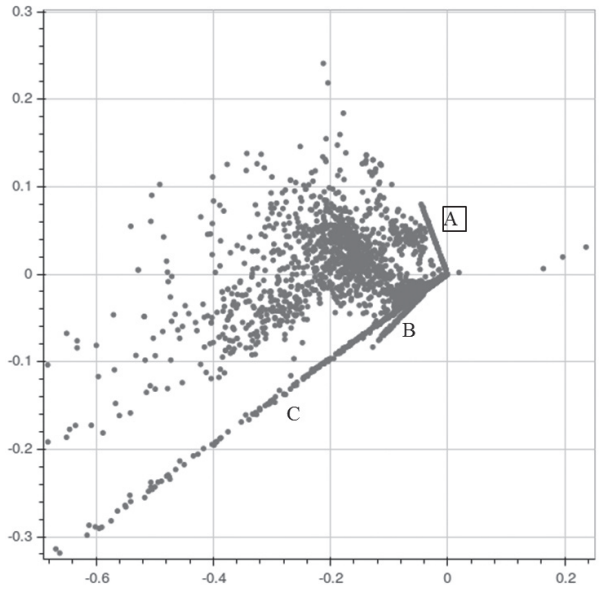
What do the axes in the graph above represent? Each axis is a certain level of each of the 141 values (dimensions), i.e., a vector as plotted below. To get back to the approximation of the original vector represented by one point in Figure 1, take the x and y coordinates, and use them to scale the x and y axis vectors in Figure 3.

Figure 3: Meaning of X and Y axes in Figure 1



Clearly, these vectors are a bit weird. There are regular dips. It turns out there are life tables with values only every several years, not every year, and those dominate the description of the data. What strikes me is that several patterns emerge anyway. Around (0,0) we have most of the vectors.

Figure 4: Zoomed in around 0,0: most (mortality, lapse, disability) vectors



Some structures jump right out. What turns out to be driving them is the areas that were missing.

A. Most points represent a vector from one of the truncated (ages 0 and 1) South American life tables.

B. Most points represent one of the South American life tables from 5-80.

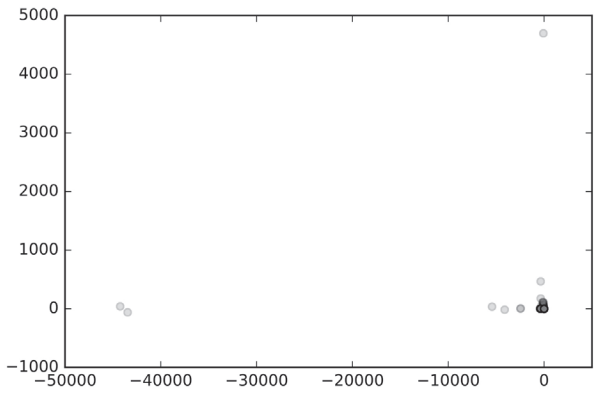
C. Most points represent disability tables or relative risk tables.

Omitting the English and Scottish life tables and others more than 200,000 from the origin (from eyeballing the graph), the

So what have we accomplished? In a quick analysis using a readily available algorithm, we've turned up an issue we can all relate to. ...

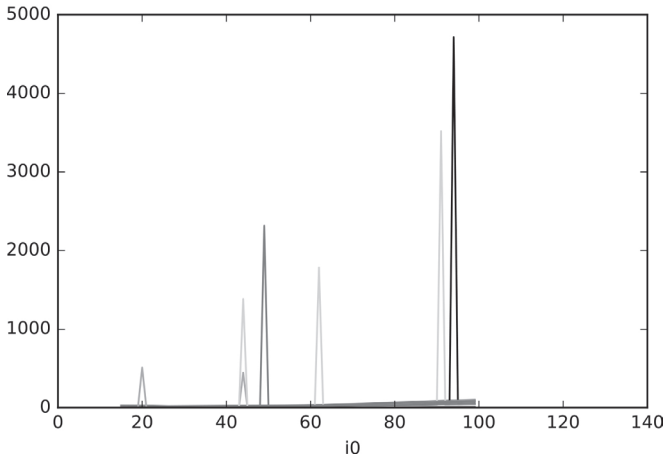
remaining tables would be plotted quite differently. There are some outliers along the y axis and some at about (-40000, 0). The former are mostly medical expense tables and the latter are more life tables. Both types are quite different from mortality rates. One of the medical expense tables is especially far off. By the way, browsing through these data I'm using a Python library called Bokeh, which allows easy browsing of large datasets. It can be told to show a text box when the mouse is over a point on the graph, which is how I tell the point's corresponding vector.

Figure 5: Decomposing again without main outliers



This outlier pointed me to some issues with scanned medical cost tables: Some values were missing decimals. The medical expense tables in question are from the 1970s and I doubt they are being used, but I'll still point it out to the table managers. That is exactly the kind of thing we are looking for. Grabbing those from the database and plotting them, we see some problems in the data entered for the 1974 medical expense tables. See Figure 6.

Figure 6: 1974 Medical Expense Tables (including typos)



Checking the values themselves, it's easy to see the decimal did not get typed or scanned for some values. To save paper, instead of printing them, I'll let you check them yourself, unless the database has been corrected by print time.

So what have we accomplished? In a quick analysis using a readily available algorithm, we've turned up an issue we can all relate to: an error in a valuation system. Are you ready to go through your own table repository?

Brian D. Holland, FSA, MAAA, is director and actuary, of Individual Life and A&H Experience Studies at AIG. He also serves as chair of the Predictive Analytics and Futurism Section Council. He can be reached at *brian.holland@aig.com.*

ENDNOTES

1 *http://mort.soa.org/ViewTable.aspx?&TableIdentity=1531.*

2 *http://mort.soa.org/ViewTable.aspx?&TableIdentity=2952.*

The Impact of Deep Learning on Investments: Exploring the implications one at a time

By Syed Danish Ali

Until recently, the artificial intelligence portion of data science was looked upon cautiously due to its history of booms and flops.[1] However, major improvements have been made in this field and now deep learning, the new leading front for AI, presents a promising prospect for overcoming big data problems.

A method of machine learning that undertakes calculations in a layered fashion, deep learning starts with high level abstractions (vision, language and other artificial intelligence-related tasks), moving to more and more specific features.[2] The machine is able to progressively learn as it digests more and more data, and its ability to transform abstract concepts into concrete realities has opened up a plethora of areas where it can be utilized. Deep learning has various architectures, such as deep neural networks, deep belief networks, deep Boltzmann machines, and so on, that are able to handle and decode complex structures with multiple nonlinear features.[3]

Deep learning offers us considerable insight into the relatively unknown, unstructured data, which is 80 percent of the data we generate, according to IBM.[4] Data analysis before 2005 focused on just the tip of the iceberg; the recent big data revolution and deep learning now offer us a better glimpse into the segment of data we know exists but are constrained in accessing. Deep learning helps us in both exploring the data and identifying connections in descriptive analytics, but these connections also help us in forecasting what the result will likely be, given the particular combination as the machine learns from the data.

Deep learning, in collaboration with other machine learning tools, is making headway in possible applications. All major giants like Google, IBM and Baidu are aggressively expanding in this direction but startups are providing the most vivid applications so far. Kensho[5] is a startup that aims to use software to perform tasks in minutes that would take analysts weeks or months. Just like searching via Google, the analysts can write their questions in the Kensho's search engine. The cloud-based software can find targeted answers to more than 65 million combinations in seconds by scanning more than 90,000 actions, which are as myriad as political events, new laws, economic reports, approval of drugs, etc., and their impact on nearly any financial instrument in the world.[6] Another startup, Ufora[7] is set to automate a large part of quantitative finance work undertaken by quants, especially on the stochastic modeling front. Even some hedge funds like Renaissance Technologies[8] are proactively working on machine learning and deep learning algorithms to better see patterns in the financial data to exploit opportunities (which stocks are overrated or underrated, when the market is going strong on fundamentals or approaching the bubble stage and so on) to guide their investment strategies.[9]

On the other hand, firms like Narrative Science[10] and Automated Insights,[11] working on text analytics, utilize deep learning to create lively and interactive narrative reports out of data and numbers. The reports—generated by a machine—read almost like they were written by a human. To elaborate, Narrative Science's Quill platform undertakes statistical analysis by applying time series, regression, etc., then the semantic engine evaluates the important data signal from the unimportant noise, per the needs of the audience in question, such as different reasoning if it is related for a quant or an investment trader. The patterns are spotted and made sense of in a holistic manner. Particular fuzzy attention is given to anomalies and elements of results that deviate from the main body of the results to ascertain their impact and proper interpretation. Quill remembers previous reports so it doesn't become repetitive. Natural language generation is applied with a surgeon's precision and expertise in forming such a dynamic semantic engine.

Deep learning allows us not just to better explore and understand the data, but also to improve forecast performance. For

predictive analytics, the startup MetaMind[12] is working to help financial firms assess the chances of selling stocks by going through corporate financial disclosures, according to its website. It identifies from previous experiences when a particular combination of actions led to a particular result to assess the chances of the same result happening in the future.

Extrapolating this trend into the future, it is my opinion that such analytics might soon find their way into mergers and acquisitions (M&A) and will be able to come up with the probability of some key event happening and the consequences of it when involved in a high stakes M&A. Another application can be to apply deep learning to help with one of the most vexing problems—financial crises. Economists, financial experts and social scientists have elaborated on a lot of key issues that lead to financial crises in general, as well as specifically for a particular meltdown. These can form the modeling methodology for the deep learning machine to analyze the cosmic scale of data available on any and every platform that it can garner. Such evaluation can perhaps help us to see patterns we may have missed otherwise as well as to allow us to understand more accurately the sequential movements and mechanisms involved in a particular financial contagion and crisis. There is no guarantee this will work. But perhaps it can shed some light inside the "quantum black box" of financial crises. This seems to be the need of the hour with recurring financial hemorrhages such as the EU crisis on Greek debt as well as the recent massive and escalating falls in Chinese stock exchanges—reminding us of the bitter past we faced in the Wall Street crisis of 2008-09.

Extrapolating this trend into the future, it is my opinion that such analytics might soon find their way into mergers and acquisitions (M&A) and will be able to come up with the probability of some key event happening and the consequences of it when involved in a high stakes M&A.

Given all these developments, there are still a myriad of issues that need clarification with not just deep learning specifically, but also with big data generally. Automation of such unprecedented scale and intensity raises the possibility of mass redundancies in the labor force across the economy. Are we comfortable with giving up our controls to such applications without knowing the full implications of such a move? Not every innovation brings positive results or sustains in the long run. Technology is progressing at an unstoppable pace, but can we manage the social consequences and make it sustainable in the long term? Human efforts are seemingly being diverted from other fields into information technology, which consequently can imply a concentration of power in one overlord field to the potential detriment of others. Are we ready for this? From a consumer point of view, how ethical is it that marketing personnel know you so well that it makes rational optimization very difficult on the part of the consumer?

These are all good questions and should be adequately and mutually tackled and addressed by all the stakeholders involved such as the data scientists, governments, professions and consumers so a mutual policy that can better alleviate such concerns can be reached. The core aim of the policy has to be to sustain technology for the benefit of our societies, to lead to value creation, to reduce scarcity and reduce fragility of our systems, as well as to generate more resources for our prosperity instead of creating the monster of Frankenstein, as "Terminator" and other doomsday movies will have us believe. ■

Syed Danish Ali is a senior consultant at SIR consultants, a leading actuarial consultancy in the Middle East and South Asia. He can be reached at *sd.ali90@ymail.com.*

ENDNOTES

1 Jack Clark, "I'll Be Back: The Return of Artificial Intelligence," Bloomberg Business, Feb. 3, 2015, *http://www.bloomberg.com/news/articles/2015-02-03/i-ll-be-back-the-return-of-artificial-intelligence.*

2 Will Knight, "Deep Learning Catches on in New Industries, From Fashion to Finance," MIT Technology Review, May 31, 2015, *https://www.technologyreview.com/s/537806/deep-learning-catches-on-in-new-industries-from-fashion-to-finance/.*

3 Yoshua Bengio, "Learning Deep Architectures for AI" (University of Montreal technical report 1312, 2009), *https://www.iro.umontreal.ca/~lisa/pointeurs/TR1312.pdf.*

4 "Analytics Overview," IBM, accessed May 17, 2016, *http://www.ibm.com/analytics/us/en/what-is-smarter-analytics/innovate-with-analytics-tools.html.*

5 *https://www.kensho.com/#/.*

6 Steven Bertoni, "Goldman Sachs Leads $15 Million Investment in Tech Start Up Kensho," Forbes, Nov. 11, 2014, *http://www.forbes.com/sites/stevenbertoni/2014/11/24/goldman-sachs-leads-15-million-investment-in-tech-start-up-kensho/#5e970f7c5795.*

7 *http://www.ufora.com/.*

8 *https://www.rentec.com/Home.action?index=true.*

9 Kelly Bit, "The $10 Hedge Fund Supercomputer That's Sweeping Wall Street," Bloomberg Business, May 20, 2015, *http://www.bloomberg.com/news/articles/2015-05-20/the-10-hedge-fund-supercomputer-that-s-sweeping-wall-street.*

10 *https://www.narrativescience.com/.*

11 *https://automatedinsights.com/.*

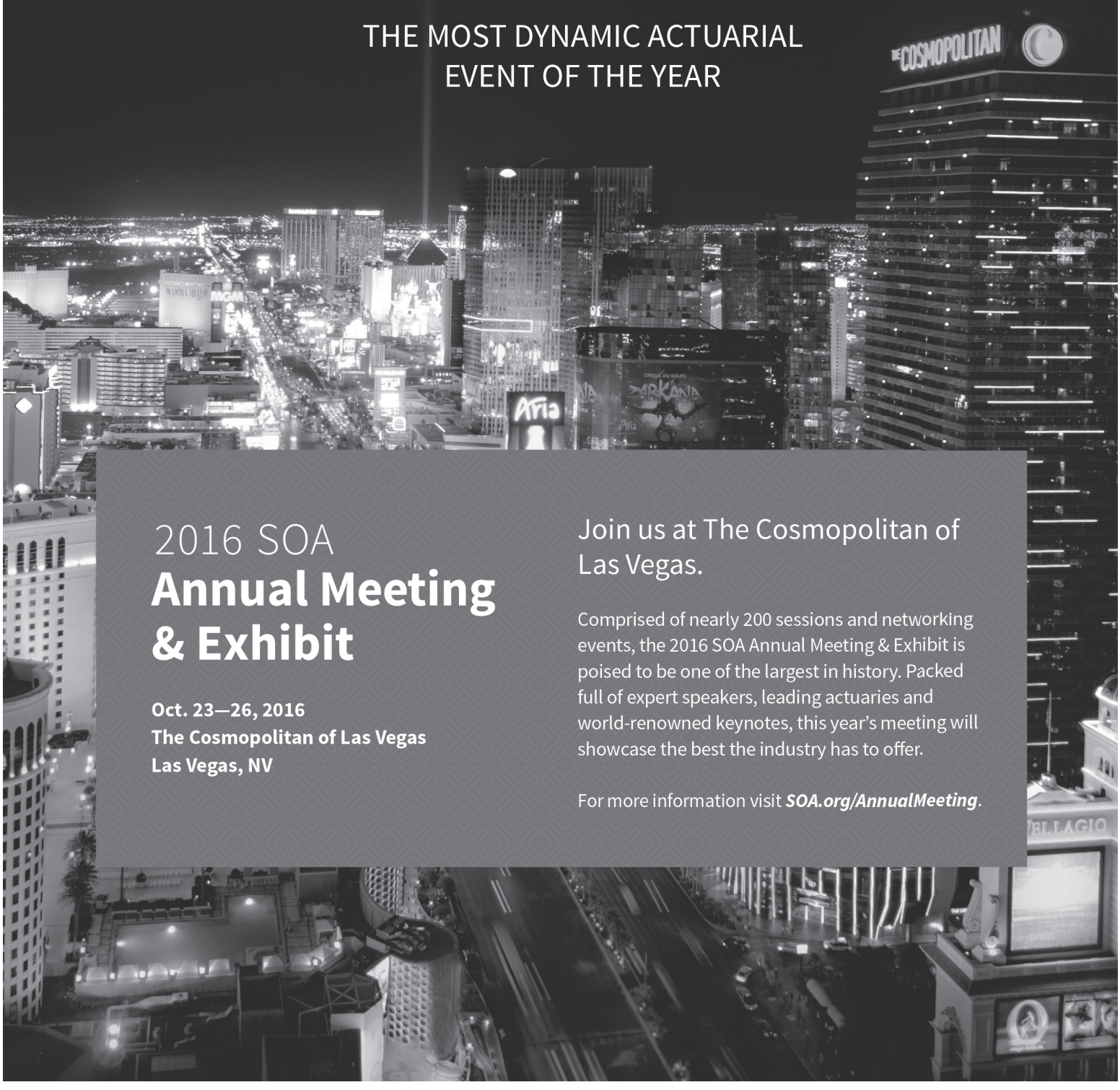
12 *https://www.metamind.io/.*

INDUSTRY INSIGHT | DIVERSE CONTENT | LATEST SOLUTIONS | EXPERT KNOWLEDGE

SOCIETY OF ACTUARIES

YOU'RE INVITED

THE MOST DYNAMIC ACTUARIAL EVENT OF THE YEAR



2016 SOA

Annual Meeting & Exhibit

Oct. 23—26, 2016
The Cosmopolitan of Las Vegas
Las Vegas, NV

Join us at The Cosmopolitan of Las Vegas.

Comprised of nearly 200 sessions and networking events, the 2016 SOA Annual Meeting & Exhibit is poised to be one of the largest in history. Packed full of expert speakers, leading actuaries and world-renowned keynotes, this year's meeting will showcase the best the industry has to offer.

For more information visit ***SOA.org/AnnualMeeting***.

SOCIETY OF ACTUARIES

475 N. Martingale Road, Suite 600
Schaumburg, Illinois 60173
p: 847.706.3500 f: 847.706.3599
w: www.soa.org

NONPROFIT ORGANIZATION
U.S. POSTAGE PAID
SAINT JOSEPH, MI
PERMIT NO. 263