SOCIETY OF ACTUARIES
PREDICTIVE ANALYTICS AND FUTURISM SECTION

Predictive Analytics and Futurism

ISSUE 19 • DECEMBER 2018

The First Step in Building a Predictive Model

By Nathan Pohle

Page 9

Predictive Analytics and Futurism

Issue 19 • December 2018

Published three times a year by the Predictive Analytics and Futurism Section of the Society of Actuaries.

475 N. Martingale Road, Suite 600
Schaumburg, Ill 60173-2226
Phone: 847.706.3500 Fax: 847.706.3599
www.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 webpage at *http://www.soa.org/predictive-analytics-and-futurism/*.

This publication is provided for informational and educational purposes only. Neither the Society of Actuaries nor the respective authors' employers make any endorsement, representation or guarantee with regard to any content, and disclaim any liability in connection with the use or misuse of any information provided herein. This publication should not be construed as professional or financial advice. Statements of fact and opinions expressed herein are those of the individual authors and are not necessarily those of the Society of Actuaries or the respective authors' employers.

Publication Schedule

Publication Month: April 2019
Articles Due: Contact editor for copy submission date

2019 SECTION LEADERSHIP

From the Editor: “And now, I are one!”

By Dave Snell

Way back in the late 1970s, Digital Equipment Corporation (DEC) and Carnegie Mellon University developed one of the world’s first commercial artificial intelligence (AI) expert systems, called R1. R1 was written in a variant of Lisp[1], and it had dozens, if not hundreds of rules. It was designed to handle the complicated and interacting constraints involved in configuring the VAX 11/780 series computers for specific customer installations. The research effort was led by John McDermott, and a running joke among the team was the comment, “I wanted to be a knowledge engineer, and now I are one.”[2] By the time I got involved in AI in the early 1980s, the line was often restated as, “Last week, I didn’t know what a knowledge engineer was, and now, I are one.”[3]

The analogy I wish to make here is that **actuaries may be about to assume a new role that we didn’t know about just a short time ago.**

This year I was privileged to chair the program committee again for the SOA Predictive Analytics Symposium. It took place in Minneapolis, Sept. 20–21, and like last year it was a resounding success. Comments from the attendees included “awesome,” “wonderful,” and “the best SOA meeting I have ever attended.”

The symposium featured six concurrent tracks of sessions on a wide variety of predictive analytics topics ranging from management perspectives through leading-edge techniques for experts. We also had keynote presentations from industry leaders, who shared their insights regarding the future of insurance opportunities. These were especially exciting. The former CEO of a reinsurer with trillions of dollars of life insurance reinsured stated that the two big areas of opportunity are AI machine learning and genetics.

It’s no secret that the American public does not hold insurance companies in high regard. Some think of actuarial work as morbid. Sure, that’s a pun deserving a groan; but many actuaries are involved in predicting morbidity and mortality, and the various financial risks associated with them (experience studies, pricing, valuation, reserves, etc.). We are bombarded with political ads for candidates who vow to stand up to the insurance companies for the rights of the downtrodden. Unfortunately, some of that animosity is understandable as the news media highlight instances of unfortunate souls being denied coverage or claims due to pre-existing medical conditions. An idea raised at our symposium was that we have an opportunity to change that perception. Two speakers noted that after issue, there is little or no follow-up with insureds to help them live longer and healthier lives. Now, with new knowledge and tools relating to machine learning and genetics, we can apply our newly acquired predictive analytics capabilities and our ability to critically evaluate clinical studies, and assimilate them with other, newer, data sources. Armed with these results, we can provide guidance for the vast insured population to enjoy life (and health) and do this in a manner that aligns our interests with those of the public. It can be a win-win relationship where we become life coaches, so to speak, who advise each person on optimal medications/procedures/lifestyle changes that are personalized for their specific mix of biological and environmental characteristics. The possibility even exists that epigenetics and precision medicine (through predictive analytics) could cure or prevent several currently uninsurable conditions and reduce the barriers to universal health care.

This issue provides several articles describing new tools, information sources, perspectives and standards for you to start assuming your new actuarial role—possibly much different from the one you used to have:

- Anders Larson begins with his “Outgoing Chairperson’s Note.” He summarizes some very impressive accomplishments of the section over the last year, and uses his son’s Magic 8-Ball to supplement the sophisticated predictive modeling tools the section has been fostering. The result is:

"Outlook Good." Please read his summary and take advantage of the many resources he describes.

- Next, we have the "Chairperson's Corner," by Eileen Burns. Eileen focuses on two specific areas where we can continue to add value, and both of them involve education. The first is education for performing predictive analytics and futurism. The second is education understanding, validating and communicating—especially to stakeholders.
- Nathan Pohle is in harmony with these goals. His article, "The First Step in Building a Predictive Model," makes the point that a best practice in the building of any model is to first focus on what problem you wish to solve. Next, build the business case. "Start with a high-level business case and an effective marketing plan to communicate the benefits of the solution internally." It's good advice!
- Ricky Trachtman, in "Ethics and Professionalism in Data Science," stresses the need to ensure that this powerful tool of predictive analytics is used in an ethical manner. We must acknowledge and address privacy issues, profiling and the implications of algorithmic decisions (perhaps with minimal human oversight) that may be based upon incomplete or inappropriate data. Ricky reminds us to "Act honestly with integrity and competence" from Precept 1 of an applicable ASOP from the American Academy of Actuaries.
- Next, Jeff Heaton gives us a "Math Test for Models." In this article, he shares his strategies for feature engineering—a critical part of any predictive analytics model (PM). Jeff shows his comparison of four common PM techniques: support vector machines, random forests, gradient boosting machines and neural networks on 10 different equation types. He provides interesting insights in the article and accompanies it with a link to the code on his GitHub site.
- Speaking of code, Alexandru Andrei describes a very handy set of code (on his GitHub site) in his article "Extracting Medical Data from Wikipedia." He wrote this to mine the vast (over 68 GB and still growing) online encyclopedia Wikipedia. His project was specifically to gather information on diseases, associated medical codes and medications; but he gives detailed instructions on how to deal with this huge dataset without running out of memory. You can modify a copy of the code for various data mining projects.
- The usual artificial neural network (ANN) has an input layer, some hidden layers (for deep learning) and an output layer. This is handy for snapshot numeric input but awkward for images or for time series analysis. Holden Duncan, in his article "The Possible Role of Convolutional Neural Networks in Mortality Risk Prediction," shows us how convolutional neural networks (CNNs) are well suited for not just images, but also data with historical information, such as mortality studies.
- Dorothy Andrews gives a useful new perspective of images. In "The Psychology of Visual Data," she describes how graphs and charts can mislead viewers into thinking numerically insignificant differences are very important. Ethical use of visuals should inform, not mislead and Dorothy provides several examples of misuses. She cites many references that can help you make better use of visuals in your presentations and proposals.
- Wrapping up this issue, Mary Pat Campbell has written a book review for us on "Actuarial Statistics with R." This textbook was written by actuaries for students preparing for the SOA and CAS actuarial statistics exams. Mary Pat gives it a good recommendation for actuarial students; but she cautions us that it is not for complete beginners to R. She offers references to other sources for just getting started. You can also refer back to her December 2015 PAF article "Getting Started in Predictive Analytics: Books and Courses" for still more sources.

Perhaps someday, the political ads will show candidates who vow to help the insurance companies help humanity. Perhaps one or more of them will be a new type of actuary. Today, I am an actuary who wants to become the new professional who applies AI and machine learning and epigenetics to help the world. Someday, I hope to say "and now, I are one!" ■

Dave Snell, FALU, FLMI, ASA, MAAA, CLU, ChFC, ARA, ACS, MCP, teaches AI Machine Learning at Maryville University in St. Louis. He can be reached at *dave@ActuariesAndTechnology.com.*

ENDNOTES

[1] No, R1 had nothing to do with the currently popular programming language R, which was created over a decade later (1992). The earliest high-level programming languages were Fortran (1950s), then Lisp(1958). Some of today's fastest R and Python packages are written in Fortran; and Lisp is still utilized in various expert systems.

[2] "R1: A rule-based configurer of computer systems," Artificial Intelligence, Volume 19, Issue 1, September 1982, pages 39–88.

[3] From 1983 to 1987, I served on the advisory board for Washington University's Center for the Study of Data Processing, in St. Louis, and we often discussed this early commercial use of AI. It encouraged us to (mistakenly) predict that AI would revolutionize computer science within a decade. McDermott grew up in St. Louis County, and came back periodically to visit and lecture.

SOCIETY OF
ACTUARIES

Outgoing Chairperson’s Note

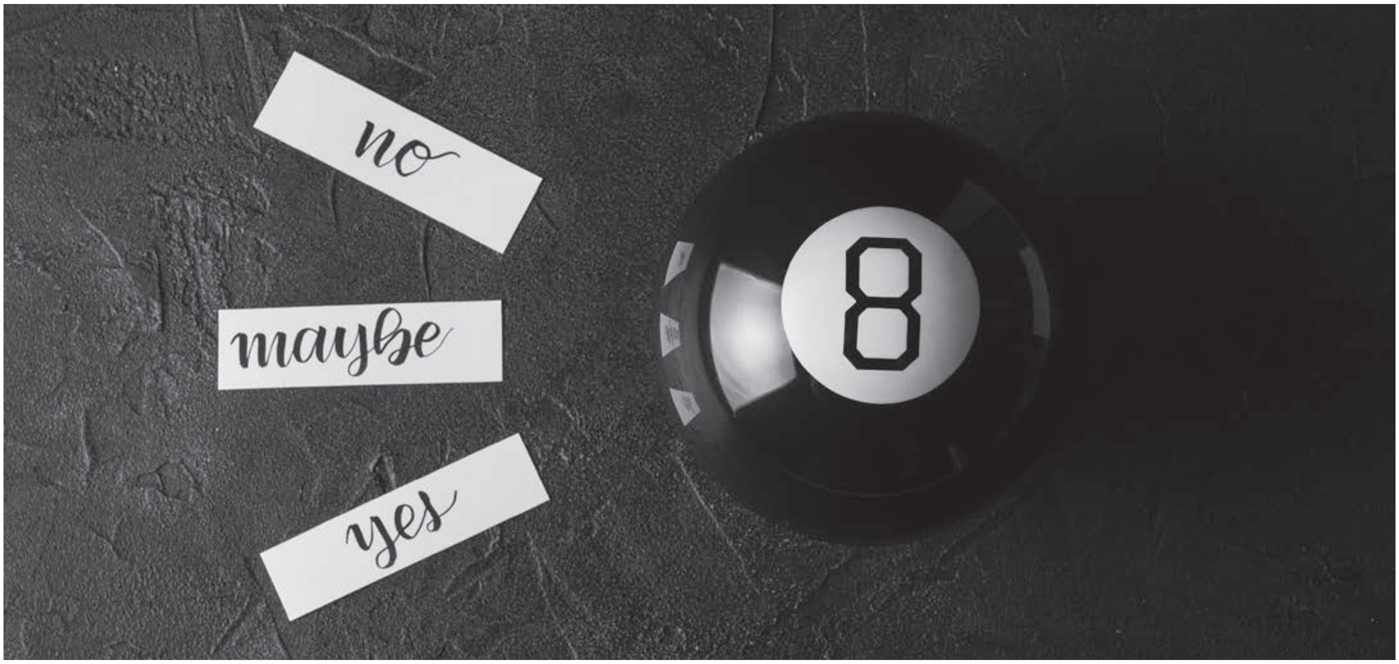
By Anders Larson

As my tenure as chairperson of the Predictive Analytics and Futurism (PAF) section was drawing to a close, I began to wonder about what’s next for the section after I leave. Predicting the future—that’s what our section is all about, right? But as I sat down to write this piece, I initially struggled to make out a clear vision of what 2019 and beyond will look like for PAF. My trusty crystal ball was failing me, and the Magic 8-Ball I recently purchased for my 4-year-old son kept vacillating between “Concentrate and ask again” and “Better not tell you now.” What a waste of $9.95.

But it struck me that predictive modeling is not just about looking forward. It’s about looking back into the past, learning from a period of time when you know the outcomes, and then applying this knowledge to make useful predictions about the future. So, let’s focus instead on what the PAF section has done over the past year. In machine learning terms, let’s review the training data. Once we’ve done that, perhaps we’ll have a better sense of what could be on the horizon.

> predictive modeling is not just about looking forward. It’s about looking back into the past …

- **We committed funding to two exciting research projects that are in progress right now**. One is a Delphi study on economic scenario generation. This project falls more on the “futurism” end of the “predictive analytics and futurism” spectrum, and we felt it was important that the section be involved in this effort. The other research project is related to validation of predictive models. The second project falls more squarely in the “predictive analytics” arena, and the section believed it was a critical area of study as more and more actuaries are incorporating machine learning and other non-traditional modeling techniques into their work.
- **Coming soon—Our own podcast feed.** Thanks in large part to the efforts of former council member Shea Parkes, the podcast has been a major success over the past three years. When we started, we were recording sporadic episodes through the office phone system. The audio quality was undeniably poor, but the content was strong, and we gained traction with SOA members. We have since upped our game by using professional microphones, and soon we will have our own feed (currently our podcasts are part of the SOA’s broader feed). Consider subscribing to our feed when it comes out and look for new episodes to be released every two months.
- **We completed a survey of current members and responded with two new initiatives.** I covered this in detail in my “Chairperson’s Corner” article in the August newsletter, but here’s the short version. We surveyed our current members in early 2018, and based on the responses from more than 250 of you, we decided to pursue two new initiatives: a Jupyter Notebook contest and a “Hack-a-thon.” We are still in the planning stages for both, but we do know that the “Hack-a-thon” will be held either at the beginning or the end of the 2019 Predictive Analytics Symposium. Look for more details on both in the coming months.
- **We seamlessly transitioned our newsletter to a three-times-per-year format**. OK, so behind the scenes, it wasn’t entirely seamless. But thanks to the hard work from council members and our editors, we were able to produce three great issues in 2018. We think this format will help us stay more current than the two-times-per-year format that had been in place for many years.
- **The 2019 Predictive Analytics Symposium built on the success of the inaugural meeting in 2018**. This meeting is certainly not entirely a PAF-led effort, but two council members were on the planning committee and many council members and friends of the council presented at this year’s meeting. Once again, enthusiasm was very high and just over 93 percent of the evaluation respondents would recommend this to others. It was gratifying to read comments such as, “This is easily my favorite SOA symposium/conference of the year! It’s full of presenters and attendees eager to push the envelope and learn more.” Another comment, “I’m looking forward to some kind of hands-on hacathony thing” was timely, as we plan to add one to the end of the symposium next year.
- **We contributed to a predictive analytics-focused issue of *The Actuary*.** Nearly every article in the June/July 2018 issue of *The Actuary* was related to predictive analytics



applications in actuarial science. While our section was not responsible for all the content, we had three articles written by council members, and council member Dorothy Andrews was co-editor of the issue along with Jacque Kirkwood of the SOA. If you haven't already done so, please take a read through this issue. You won't be disappointed.

- **We kept doing all the other stuff you've come to expect from us, which is no small feat**. Every year, the section puts on the Practical Predictive Analytics Seminar after the Life and Annuity Symposium, sponsors several sessions at each of the major SOA meetings, helps organize webcasts, maintains a social media presence on LinkedIn, and more. Although these aren't new initiatives, they require time and energy from council members, friends of the council, and our partners at the SOA. Thanks to all who volunteered with these efforts during 2018!

OK, so what does the "data" tell us about what the future might hold for the PAF section? I think what it tells us is that creative, energetic volunteers can help drive some exciting outcomes. And what I do know is we have plenty of those on the council and as friends of the council. I also know that our new chairperson, Eileen Burns, brings a wealth of predictive analytics knowledge and experience to the leadership role. With that said, I can't help but conclude that the PAF section has plenty more good things to come in 2019 and beyond.

But just for good measure, let's consult the Magic 8-Ball one more time. "What can you tell us about the PAF section?"

[Shakes ball] "Outlook good."

Now we're talking. ■

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

Chairperson's Corner

By Eileen Burns

As the incoming chair of our section, I'd like to take this opportunity to welcome our new council members, Xiaojie (Jane) Wang, Michael Niemerg and Garfield Francis! We're looking forward to leveraging your interest and enthusiasm in predictive analytics and futurism to advance knowledge among actuaries in this area. I'd also like to offer my thoughts on two main areas in which I hope our section will continue to raise the bar in predictive analytics and futurism in the coming year.

First, I hope and expect the Predictive Analytics and Futurism (PAF) Section to continue to deliver on our educational promises. Our section's purpose, stated front and center on its section site on SOA.org, is "to examine the advanced methods and tools used in predictive analytics and futurism through professional development, meetings, special seminars, research studies and the creation and dissemination of literature."

> I hope and expect the Predictive Analytics and Futurism Section to continue to deliver on our educational promises.

Anders did a wonderful job recapping the many activities we've completed under this umbrella in the past year—sponsoring sessions at five SOA meetings, sponsoring the Practical Predictive Analytics Seminar, a transition from two to three newsletters, establishing our own podcast feed, helping produce a predictive analytics-focused issue of *The Actuary*, and starting two research endeavors—and foreshadowing two more activities under development: A Jupyter Notebook contest (first) and a Hack-a-thon (to follow).

I could almost say, based on our past record, that we're done—we've met our purpose! But the fact is that predictive analytics is an emerging area of expertise, with many important questions left to be answered, and best practices to be agreed.

There are two specific areas where we as a section can continue to add value:

1. Providing education on existing and emerging methods for performing predictive analytics and futurism.
2. Providing education related to understanding, validating and communicating predictive analytics to stakeholders.

There will continue to be new methods for predictive analytics until all questions have been answered about the future, so I expect for years to come we'll continue to be able to offer podcasts on new data-cleaning methods, new modeling methods, new tools for visualization and validation, and the like.

In particular, as I wrote in our last newsletter, we still have a way to go in ensuring stakeholders understand and trust predictive models. While the next year or two is likely to see a lot of movement in this area, as long as modeling methods are evolving so too will methods for model validation. I expect education in methodology and validation to be ongoing endeavors as long as we are a section.

Second, I hope that we will build on our collaborative efforts with other sections and with the SOA. There are often times when our interests are aligned with other sections, such as Technology (how do you use the cloud?), Actuary of the Future (e.g., what jobs should I or my team prepare for?), or Entrepreneurship and Innovation (e.g., how can I use these mathematical models to provide a new actuarial service?). I hope to build on the relationships among our sections to ensure that the full breadth of relevant topics can be represented in sessions at the major SOA meetings, in newsletters, and in webcasts.

In the broader educational sense, the SOA is working to provide educational offerings that will keep actuaries relevant for employment openings in the area of predictive analytics. As practitioners in this area, members of our section have unique insights into what is required now, and what will be required in the future. I hope that we will find opportunities to work with the SOA to define and create content related to this important topic.

Sometimes, a Delphi study (this is from our futurism side if you weren't aware) can point you to the right question to ask. Clearly Anders's thought experiment led him to the right question for his Magic 8-Ball, "What can you tell us about the PAF Section?"

"Outlook good." ■

Eileen S. Burns, FSA, MAAA, is a principal and consulting actuary with Milliman. She can be contacted at *eileen.burns@milliman.com*.

The First Step in Building a Predictive Model

By Nathan Pohle

There is a lot of excitement in the actuarial community about the potential of machine learning and predictive models. As a result, a lot of the conversation has focused on the technical details, including the types of Machine Learning (e.g., supervised vs. unsupervised), the types of programming language (e.g., R or Python), and other modeling details. While the excitement around predictive modeling and the technical details is merited and the discussion of these details is important, there are other considerations that should be included in the discussion. This article covers some of the broader considerations for any predictive model, and for that matter, any model or project.

A good starting point for any model is to define what question(s) you are trying to answer. What are you trying to solve for and what is the outcome you are trying to achieve? It can be easy to be proud of the technical aspects and details of a given model, and at the same time lose track of the big picture. Creating a North Star for the goals and objectives of a model before starting to write the first line of code will help ensure the model is fit for purpose.

In order to achieve that North Star, it is imperative to start collaborating with other functions and workforce segments before the model build begins. For example, if the model being developed will impact marketing, distribution and operations, it is important to work with these departments to establish buy-in from "Day 1." That way, these departments can have a voice in what needs to be considered in the model, along with having the advanced notice to prepare for any changes that the modeling project will have on their people, processes and/or systems. Likewise, those other departments can inform you of any current or planned initiatives that may impact your project.

Once the questions, goals and objectives are landed upon, the next step should be to effectively build the business case for that model or project. This is an area that actuaries are uniquely trained to perform, given their technical acumen, product/business knowledge and risk awareness. The business case should include a high-level description of the project/model, the estimated impact (e.g., revenue gain, margin gain, cost savings), resource needs for the project, and any risks and interdependencies. Building a business case will help crystalize whether it is worth the time and effort to build the model. It will help crystallize the North Star, which is the outcome that the model is trying to achieve.

The high-level business case is critical for further collaboration and buy-in from key stakeholders. Without buy-in from internal stakeholders, frequently a model is not worth building, as it likely won't be used. When advocating with stakeholders, do not underestimate the importance of marketing and branding. A common pitfall is to use too much technical jargon—keep the messaging at the level that senior stakeholders care about. Put yourselves in the stakeholder's "shoes" to craft your message. The business case can be a guide, as if the message doesn't translate to the metrics summarized in the business case, the senior stakeholder may not fully support. Make sure to cover both the upside and the downside risks, such as the opportunity costs of not acting and the impact on the organization. Focus on the "why" and "so what," rather than the "what."

A common pitfall is to underestimate the importance of collaboration with other functions from the beginning and appropriately marketing/branding the idea internally. Underinvesting or any mistiming can cause issues with the modeling project. Therefore, when beginning the process, it is important not to start with the details of the type of model, as senior leadership likely doesn't care as much about those details. Start with a high-level business case and an effective marketing plan to communicate the benefits of the solution internally. Focus on the high-level impacts to the business, as those are important to senior leadership. Then, after stakeholder buy-in has been achieved through an effective business case, the actuary can focus on the granular aspects of whether it is a neural network or a boosted tree algorithm. ■

Nathan Pohle, FSA, CERA, MAAA, is a consulting actuary with experience in the life insurance and sports industries. He can be reached at *npohle@deloitte.com.*

Ethics and Professionalism in Data Science

By Ricky Trachtman

The use of predictive analytics, big data, machine learning, artificial intelligence, etc., is sweeping through every aspect of our lives in one way or another. Advancements in these fields have been fast and continuous. As new techniques and utilization of big data has become more prevalent, questions about the appropriateness of how and where this information and techniques should be used have arisen. These questions have made the public and institutions think about ethical usage of the technology and data for all industries.

Nowadays it's not hard to find examples of unethical issues in predictive analytics and big data highlighted in the media. It can really make one question how fast-moving technology can have ramifications and effects not fully considered when initially developed. Many of these examples involve privacy issues, profiling of individuals and discrimination.

A good example of privacy issues was highlighted in the case of a father who finds out that his teenage daughter is pregnant thanks to targeted advertising by Target.[1] In this case, the company had started targeted advertising towards pregnant women. The father of the teenager went to the local Target to complain about his daughter receiving these ads, just to call some time later to apologize after learning that his daughter was indeed pregnant.

An example of profiling of individuals and discrimination can be found with the risk assessments produced by software being used across the country in the criminal justice system.[2] This software is used to predict "future criminals" and has been shown to be biased against ethnic groups.

In her book, "Weapons of Math Destruction," Cathy O'Neil[3] describes many ways data analytics can damage peoples' lives and further increase inequality of certain groups of people. She highlights how using predictive models can further marginalize certain populations by using proxies for data that can't directly be measured: For example the use of zip codes as a proxy for race or income. It is an interesting read that demonstrates emerging issues related to the use of predictive models via real-life examples. In her examples, O'Neil describes how some models can become less transparent and seem like black boxes. This lack of transparency contributes to the inability of individuals to know what data was used to come up with the responses the model's algorithms may produce. It is clear that the lack of transparency offered by some models can potentially hide ethically questionable practices.

After reading many of these examples and O'Neil's book, I started to think about the ramifications predictive models can have. I wanted to learn more about the root of these problems. While doing so, I stumbled across a good study which helps frame the ethical issue problems in a slightly different way. In the study, "Ethical Implications of Big Data Analytics"[4], which is a Delphi study of academic and practitioner experts, the authors categorize big data analytics as a social process. By viewing big data analytics as a social process, the authors segment the different stakeholders into three groups: individuals, organizations and society. This allows one to indicate differences in the roles each of these stakeholders have in the overall process of big data analytics.

The social process of big data analytics is described as follows. Individuals contribute data by using social media, using digital devices (internet-of-things), and having transactions with different organizations. Organizations then try to identify patterns and relationships in the data with monetary goals. The data from the individuals is also shared and sold to data brokers or other aggregators that parse information from possibly many sources. From the societal perspective, big data analytics has introduced new markets and new practices from distinct companies and entities. These innovations and new markets may require some form of regulation in order to protect and maintain fairness amongst all members of society.

Throughout this process, there are many areas in which ethical issues applied to each of the stakeholders can arise. The study goes on and identifies ethical issues associated with the three distinct stakeholder groups. Some of the issues mentioned are privacy issues, algorithmic decision-making with less than optimal data as opposed to human decision-making responsibilities, profiling of individuals by grouping them into cohorts, control and surveillance of individuals by limiting their choices and obtaining their data, and lack of transparency in the analytics value chain since there is an asymmetrical knowledge by different stakeholders.

As the world of predictive analytics and big data invades the insurance industry, it's easy to see how the issues highlighted in the Delphi study just mentioned come into play. This is where being an actuary may be an advantage against many other data analytics practitioners. As members of a U.S. actuarial organization we are bound by well-defined standards of practices (ASOPs) and a professional code of conduct (Code), which we need to uphold.

The American Academy of Actuaries' Big Data Task Force recently released a monograph "Big Data and the Role of the Actuary"[5] that discusses ethical and professional issues for actuaries working within the analytics fields. Every actuary who works or wants to work with predictive analytics should read this document. This document contains a discussion of the effects that big data have on the consumers (individuals), insurers (organizations) and regulators (society). It also provides considerations for actuaries in the use of predictive analytics and data sources. It contains a section on regulatory considerations outlining the benefits and challenges to different stakeholders, framing the existing regulatory environment as well as emerging regulatory developments. The document ends with a section on professionalism, highlighting the code of conduct precepts that are relevant for working with big data analytics and outlines the ASOPs that are applicable.

This monograph acknowledges that when there is demand for new skills evolving from the traditional actuarial practices, a wide range of ethical and professional challenges may emerge. In other words, it provides a reminder that the Code and the ASOPs can guide us through these emerging challenges while we as actuaries continue to acquire new skill sets, allowing us to keep up with emerging technologies. It also remind us of the "look in the mirror" test that is implied in the Code, which means that we should always think about our qualifications based on education and experience and make a judgment about whether we can fulfill our obligations under the code to:

- *"Act honestly, with integrity and competence—perform actuarial services with skill and care (Precept 1); and*
- *Perform actuarial services only when qualified to do so (Precept 2)."*

Predictive analytics will continue to evolve with new applications and techniques appearing rapidly in all aspects of our lives. Working at the boundaries of big data analytics will continue to shed light into potential ethical issues that we may not even think are a possibility today. Regardless of this ever-shifting environment I rest assured that we, actuaries, will continue to be at the forefront acting in a professional and ethical way. ■

Ricky Trachtman, FSA, MAAA, is principal and consulting actuary at Milliman in Buffalo Grove, Ill. He can be reached at *ricardo.trachtman@milliman.com.*

ENDNOTES

[1] Duhig, C. 2012. "How Companies Learn Your Secrets?," New York Times (available at *http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html?_r=0).*

[2] Angwin, Julia & Larson, Jeff & Mattu, Surya & Kirchner, Lauren. 2016. "Machine Bias, There's software used across the country to predict future criminals. And it's biased against blacks." ProPublica. (available at *https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing*)

[3] O'Neil, Cathy. 2016. "Weapons of Math Destruction, How big data increases inequality and threatens democracy", Crown Books. ISBN 0553418815.

[4] Asadi Someh, Ida & Breidbach, Christoph & Shanks, Graeme & Davern, Michael. (2016). ETHICAL IMPLICATIONS OF BIG DATA ANALYTICS. (available at *https://www.researchgate.net/publication/308024119_ETHICAL_IMPLICATIONS_OF_BIG_DATA_ANALYTICS*)

[5] American Academy of Actuaries Big Data Task Force. 2018 "Big Data and the Role of the Actuary" (available at *http://www.actuary.org/files/publications/BigDataAndTheRoleOfTheActuary.pdf*)

A Math Test for Models

By Jeff Heaton

In my opinion, feature engineering is one of the most critical components of any predictive analytics project. Feature engineering is a preprocessing step where additional features, or input columns, are calculated to augment or replace the original features. This technique is closely related to the similar strategies of interaction terms and column transformations.

Feature engineering is often cited as one of the most important components in many winning Kaggle competition entries. There have been many attempts to automate feature engineering. Techniques such as auto encoders, deep feature synthesis and various dimensionality reduction algorithms can provide new features that provide additional lift to models. However, feature engineering remains one of the key areas where a human data scientist excels over their mechanical automated machine learning counterparts. I suspect that when this tide shifts, we will see Kaggle leaderboards dominated by automatic machine learning entries.

MY STRATEGIES FOR FEATURE ENGINEERING

At the highest level, there are two types of features that can be engineered. The first type are simple transformations and interactions between the existing features. For this type, you might take the log of one of the features or divide one feature by another. All of the information needed to create these features is contained entirely within the data set itself. The second type of engineered feature is an augmentation. For this feature you tap external data sources to bring more meaning to features you already have. Consider a column that contains applicant's zip codes. Alone, a zip code is difficult to use in a model. However, you might use a table that contains the coordinates of the center of these zip codes to calculate distance to a major metropolitan area.

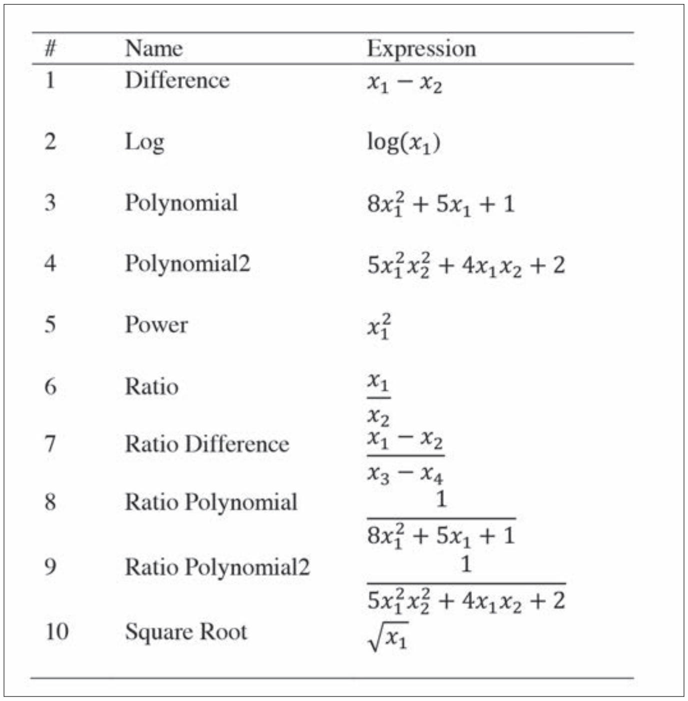
The first type of engineered feature is the focus of this article. We will examine what types of transformations and interactions will be the most effective for a variety of models. Some model types have the ability to automatically incorporate interactions. Neural networks and tree-based models in particular have this capability. I published a paper with the IEEE that explored the effectiveness of various models at self-engineering certain types of features on their own. For example, if a given model is often effective at engineering a particular form of feature, you will probably not increase the lift of that model by adding that feature type. By feature types I mean ratios, power transformations, log transformations and others.

To explore the automatic feature engineering capabilities of these model types, I created a "math test for models." I selected four model types and 10 different equation formats. For each of these equation types, I generated training data where the outcome was the result of the given equation. No noise was used. I was interested only in how well the given model could approximate the selected equation type. The results showed two interesting results. The first is that some models were more effective at certain equation types than others. The second is that certain equations are indeed much more difficult for these models to approximate. This can serve as a guide to the structure of engineered features to consider for a particular model type. The models examined in this research were: neural networks, support vector machines, random forests and gradient boosting machines (GBM). The generalized linear model (GLM) family was not considered due to their inability to automatically express interactions and transformations.

DESIGNING A MATH TEST FOR MODELS

To evaluate the automatic feature engineering effectiveness of the four model types a total of 10 different equation forms were used. The models were tested on their ability to approximate these 10 functions. If the model can easily approximate a function then it can probably automatically engineer a similar feature. The 10 different equation types are provided in Table 1.

Table 1
Ten Different Equation Types



#	Name	Expression
1	Difference	$x_1 - x_2$
2	Log	$\log(x_1)$
3	Polynomial	$8x_1^2 + 5x_1 + 1$
4	Polynomial2	$5x_1^2x_2^2 + 4x_1x_2 + 2$
5	Power	x_1^2
6	Ratio	$\frac{x_1}{x_2}$
7	Ratio Difference	$\frac{x_1 - x_2}{x_3 - x_4}$
8	Ratio Polynomial	$\frac{1}{8x_1^2 + 5x_1 + 1}$
9	Ratio Polynomial2	$\frac{1}{5x_1^2x_2^2 + 4x_1x_2 + 2}$
10	Square Root	$\sqrt{x_1}$

Simple transformations, such as log, contain only one value for x, whereas expressions like difference and polynomial contain two. The most complex equation, the ratio difference, contains four x values.

RESULTS OF THE TEST

The first step is to generate the training data. This is done by uniformly generating 10,000 random inputs for all of the x values of each equation. The y values are the result of each equation. There is no noise generated.

The math test was conducted by running each of the 10 equations against each of the four model types a total of five cycles. This results in 200 total runs. This entire process takes approximately 30 minutes to complete. The complete source code for this experiment can be found at the author's GitHub repository.[1] The five cycles are due to the fact that some of these models make use of random numbers in their training. The best (lowest) root mean square error (RMSE) score of all five cycles for each model and equation type are given by Table 2.

Table 2
Best RMSE Scores

	SVM	RF	GBM	Neural
Diff	0.03	0.00	0.01	0.00
Log	0.09	0.00	0.00	0.01
Poly	0.06	0.00	0.00	0.01
Poly2	0.05	0.02	0.02	0.01
Power	0.07	0.01	0.01	0.07
Ratio	0.66	0.14	0.21	0.15
R.Diff	28.57	100.20	204.40	28.27
R.Poly	0.03	0.00	0.00	0.00
R.Poly2	0.06	0.00	0.00	0.00
Sqrt	0.10	0.00	0.00	0.01

All errors are measured in RMSE. I did consider normalizing the output of these equations. While the domain of each equation is intentionally set to between -10 and +10, the range varies depending on the equation. Some of the equations have much larger ranges than others. A common normalization in this case is to divide the RMSE by either the mean or difference of the maximum and minimum y values. Because RMSE is in the same units as the y-value, a function with a large range will likely always have a larger RMSE than one with a small range.

I decided not to normalize because I care how closely the model approximates the function. The actual range of the function does not matter. I simply care how close the approximation is. If the approximation is perfect then the RMSE should approach zero, regardless of how large the range is.

This is shown graphically by Figures 1, 2, 3 and 4. The taller bars in each graph indicate a particular equation type that is more difficult for a given model to approximate. As can be seen from the figures, all of the models had difficulty with the ratio of differences (equation 7). The ratio (equation 6) was impossible for the support vector machine (SVM), but only somewhat more difficult for the other three model types. All other equation types were trivial for the various models to approximate.

Figure 1
Neural Network Results

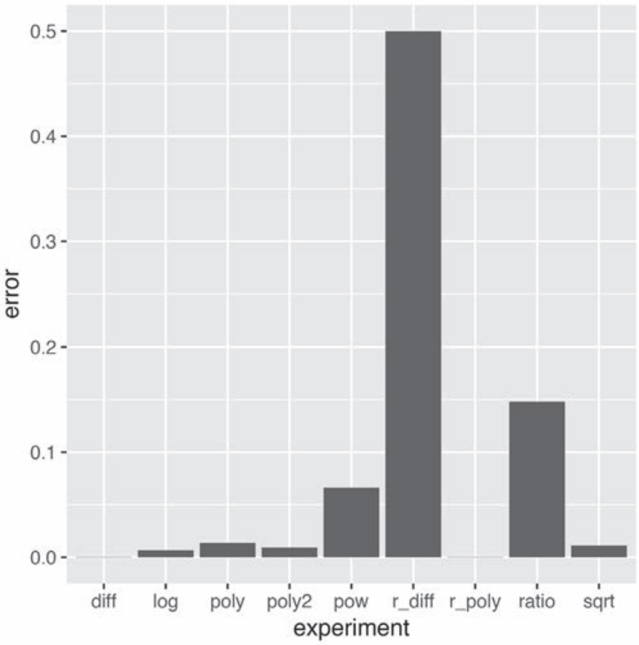


Figure 3
Random Forest Results

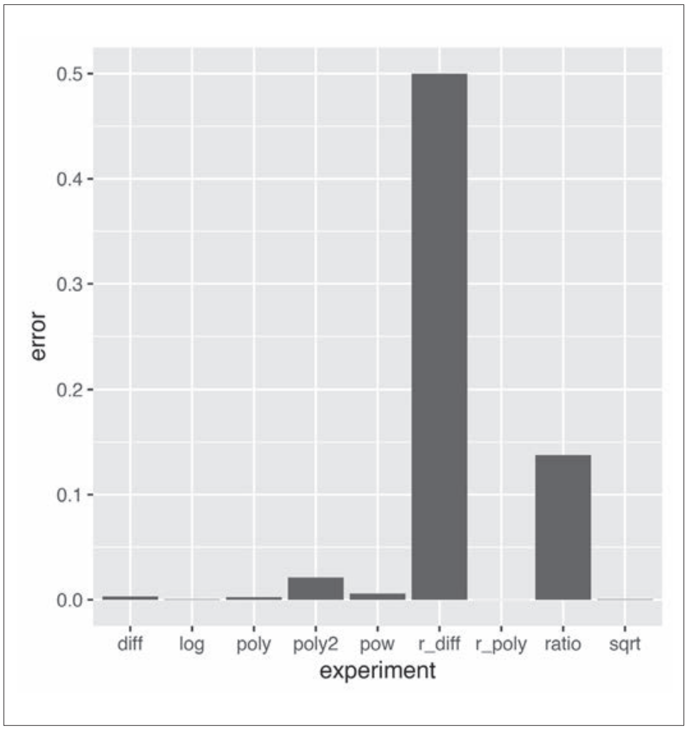


Figure 2
Gradient Boosting Machine (GBM) Results

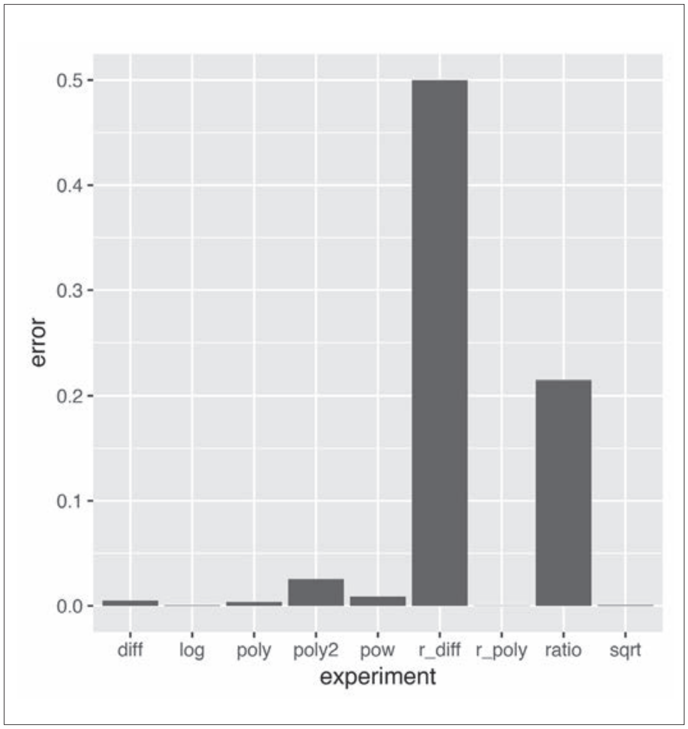
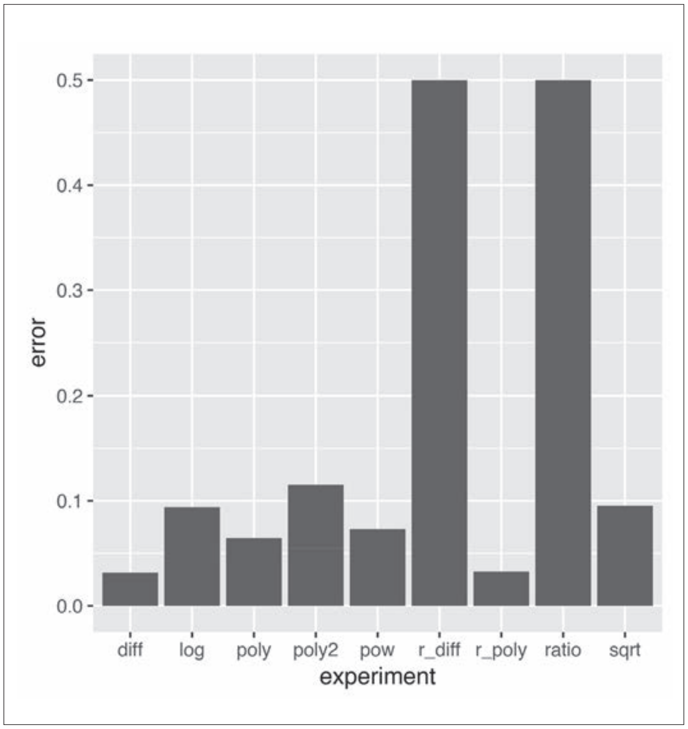


Figure 4
Support Vector Machine (SVM) Results



CONCLUSIONS

The ratio of differences was found to be a difficult feature for all models. When I am engineering my own features I frequently use ratios or a ratio of differences. The ratio by itself is a normalizer. Adding the difference causes it to be a normalizer with thresholds. For example, an engineered feature that I recently created is as follows:

(Home price – mean home price) / (age – mean age)

This is the ratio of two differences, so it has the potential to help any of the four model types. This essentially looks at how much above or below an individual's house is from the mean. However, this is then normalized by how old the individual is relative to their zip code. This reflects the fact that older individuals typically have more expensive houses than younger. Adding this calculated feature provided lift to my model.

As I continue this line of research I will introduce additional equation types to see which ones prove the most challenging for each model. I will also look at how the models might be augmented to perhaps have a chance of engineering a feature of this form. ■

Jeff Heaton, Ph.D. is VP, Data Science, RGA Reinsurance Company, in Chesterfield, Mo. He can be reached at *jheaton@rgare.com.*

ENDNOTE

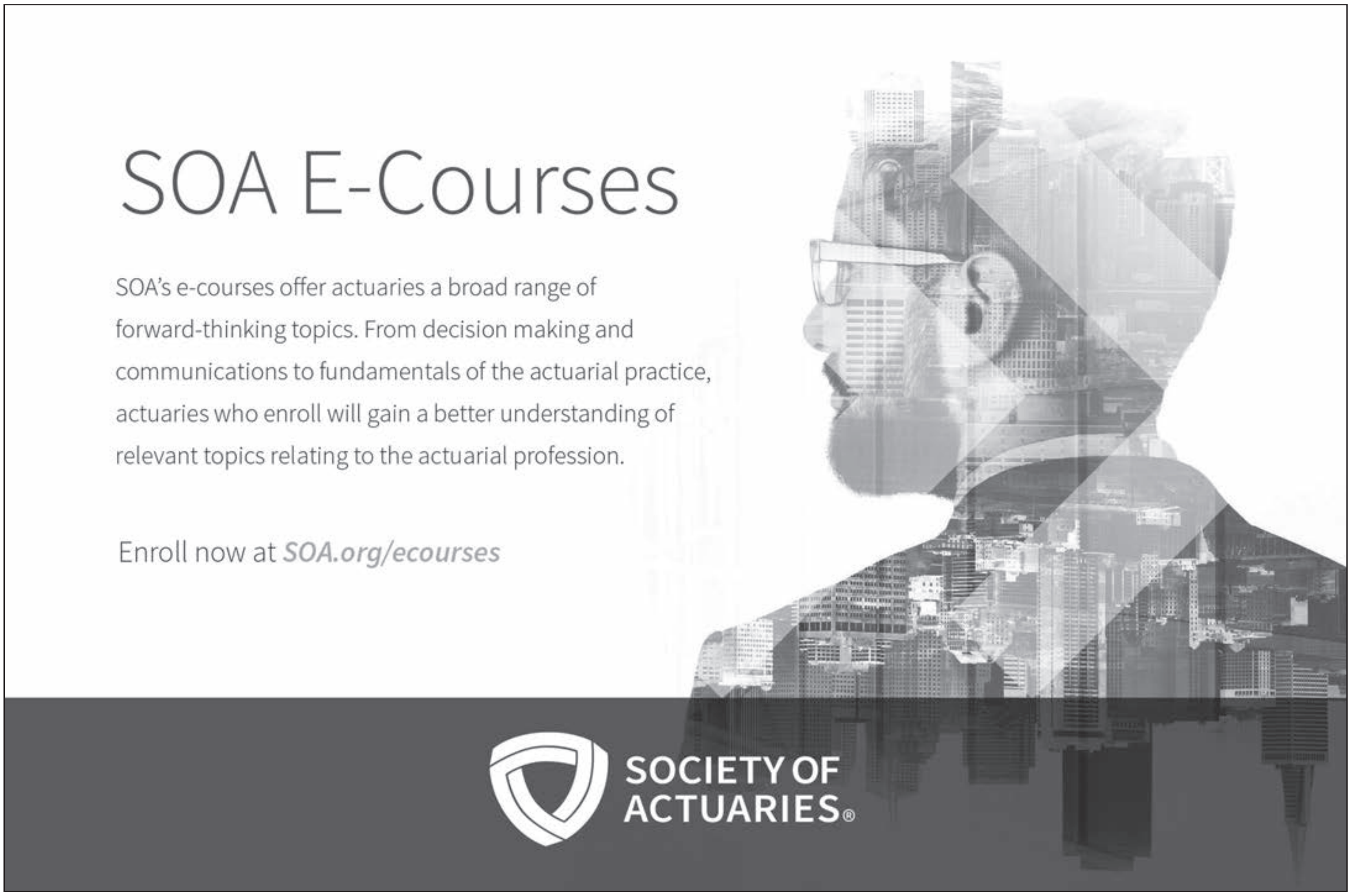
[1] *https://github.com/jeffheaton/present/tree/master/SOA/paf-mathtest*

REFERENCES

Heaton, J. (2016, March). An empirical analysis of feature engineering for predictive modeling. In SoutheastCon 2016 (pp. 1-6). IEEE.

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., ... & Kudlur, M. (2016, November). Tensorflow: a system for large-scale machine learning. In OSDI (Vol. 16, pp. 265-283).

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. *Journal of machine learning research,* 12(Oct), 2825-2830.



Extracting Medical Data From Wikipedia

By Alexandru L. Andrei

Information that includes financial, medical, demographic data, as well as facts about natural disasters is of great interest to the insurance industry. As a result, major companies, including IBM, have trained their natural language processing models using Wikipedia's dataset, one of the most comprehensive sources of information on the Internet. Any specific piece of information can be extracted from such a massive database, but the main challenge lies in processing such an enormous dataset.

Wikipedia's data is written in an XML (Extensible Markup Language) format and is presented in a manner that is both human and machine readable. However, using Wikipedia API to extract useful information is impractical due to time and traffic constraints. The file used in this article **enwiki-latest-pages-articles.xml**[1] is approximately 12.8 GB when compressed and 68 GB when decompressed, thus the majority of text editors are not able to open it. This paper describes techniques that can be used to access any specific piece of information in Wikipedia, in particular medical information, by using Wikipedia's most up-to-date dump file.

EXPLORING WIKIPEDIA DATA

The Wikipedia file is coded in the following format:

```
<mediawiki xmlns=”http://www.mediawiki.org/xml/
export-0.10/” xmlns:xsi=”http://www.w3.org/2001/
XMLSchema-instance” xsi:schemaLocation=”http://www.
mediawiki.org/xml/export-0.10/ http://www.mrt-0.10.
xsd” version=”0.10” xml:lang=”en”>
  <siteinfo>
    <sitename>Wikipedia</sitename>
    <dbname>enwiki</dbname>
    <base>https://en.wikipedia.org/wiki/Main_Page</
     base>
    <generator>MediaWiki 1.32.0-wmf.14</generator>
    <case>first-letter</case>
    <namespaces>
      <namespace key=”-2” case=”first-letter”>Media</
      namespace>
      ...
      <namespace key=”2303” case=”case-sensitive”>
      Gadget definition talk</namespace>
    </namespaces>
  </siteinfo>
. . .
  <page>
      <page>
    <title>Anarchism</title>
    <ns>0</ns>
    <id>12</id>
    <revision>
      <id>851684166</id>
      <parentid>851684072</parentid>
      <timestamp>2018-07-23T22:38:25Z</timestamp>
      <contributor>
        <username>Tajotep</username>
        <id>29695403</id>
      </contributor>
      <comment>/* Individualist anarchism */</
       comment>
      <model>wikitext</model>
      <format>text/x-wiki</format>
      <text xml:space=”preserve”>{{Use dmy dates|
       date=July 2018}}
{{redirect2|Anarchist|Anarchists|the fictional
character|Anarchist (comics)|other uses|Anarchists
(disambiguation)}}
{{pp-move-indef}}
{{use British English|date=January 2014}}
{{Anarchism sidebar
}}
{{Basic forms of government}}
. . .
. . .
      <sha1>9o00w06nsedf733vnmy7al780ia633d</sha1>
    </revision>
  </page>
```

Handling overly large files has been a major challenge of this project.

Each page within the format of the XML dump file has the following tags:

Title: contains the title of the article,
Id: the internal id of the article,
Redirect: what the page redirects to,
Namespace: helps identify the kind of page it is, and
Text: contains the information of the article itself.

HANDLING LARGE FILES

Handling overly large files has been a major challenge of this project. Regular Python parsing, which would require at least 70GB of available memory to load the XML file, is not feasible. Using the ElementTree package will allow us to load the XML file piece-by-piece without running into any memory issues. Being able to clear an element that was just processed frees memory space, allowing the next element to be loaded for processing.

```
import xml.etree.ElementTree as etree
[..]
for event, elem in etree.iterparse(pathWikiXML,
events=(‘start’, ‘end’)):
    tname = strip_tag_name(elem.tag)
     ...
    elem.clear()
```

WIKIPEDIA TEMPLATES

Wikipedia uses Wiki Markup language to create all of its articles. The focus of each search is based on a specific medical condition. The following is a template that can be used to extract specific pieces of information including a condition’s commonly prescribed medications, prognosis and mortality statistics:

```
{{Infobox medical condition (new)
| name            = <!--{{PAGENAME}} by default-->
 . . .
| medication       =
| prognosis        =
| frequency        =
| deaths           =
}}
```

Furthermore, the following template can explore a specific drug in relation to its brand and genericname as well as to its chemical properties:

```
{{Infobox drug
| drug_name         =
| INN               =
| type              =<!-- empty -->
| IUPAC_name        =
| image             =
| alt               =
| caption           =
<!-- Clinical data -->
| pronounce         =
. . .
| legal_status      = <!-- Free text -->
<!-- Pharmacokinetic data -->
| bioavailability   =
| protein_bound     =
. . .
| excretion         =
<!-- Identifiers -->
| CAS_number        =
. . .
| PubChem           =
| UNII              =
| DrugBank          =
<!-- Chemical and physical data -->
| chemical_formula  =
| molecular_weight  =
}}
```

MINING DRUG INFORMATION

The following template format enables the user to parse drug information data in a systematic way. Using string manipulation, for example, any page that contains the Drugbox template can be detected:

```
def get_drugbox(s):
    beg = (s.rfind(‘{{Drugbox’))
    end  =(s.rfind(‘\n}}’))
    if( end == -1):
        end = end =(s.rfind(‘}}\n’))
    if( end == -1):
        end = end =(s.rfind(‘}}\n<!--’))
    if( end == -1):
        end = end =(s.rfind(‘}}\n==’))
    if( end == -1):
        end = end =(s.rfind(‘}}\n\d’))
    s = s[beg: end+2]
```

After retrieving the Drugbox information, all of the data can be successfully mined. A range of codes can then be extracted from drug profiles such as CAS number, ATC and unique ingredient identifier (UNII) in addition to its chemical composition, most of which are unique to each English-speaking country.

For this project, the UNII code is of particular interest. The UNII is a unique, non-proprietary, free, unambiguous, non-semantic, alphanumeric identifier linked to a substance's molecular structure or descriptive information by the Substance Registration System (SRS) of the Food and Drug Administration (FDA) and the United States Pharmacopeia (USP). Below is the code required to extract the UNII code from the Drugbox template:

```
def find_unii(s):
    s = re.findall(r’UNII\s*?=\s?.*’,s)
    if(len(s)>0):
        s = s[0]
        equal = s.rfind(‘=’)
        #if there is a space after the equal
        remove it
        if(s[equal+1]==’ ‘):
            s = s[equal+2:]
        else:
            s = s[equal+1:]
    else:
        s = ‘’
    return s
```

MINING SPECIFIC INFORMATION FOR EACH DISEASE

Due to the fact that the data in Wikipedia is typed manually by millions of individuals, various spacing and formats must be identified. By using the “InfoBox medical condition template” below, information about each disease can be extracted:

```
def get_med_cond(s):
    beg = (s.rfind(‘{{Infobox medical
    condition’))
    end  =(s.rfind(‘\n}}’))
    if( end == -1):
        end = end =(s.rfind(‘}}\n’))
    if( end == -1):
        end = end =(s.rfind(‘}}\n<!--’))
    if( end == -1):
        end = end =(s.rfind(‘}}\n==’))
    if( end == -1):
        end = end =(s.rfind(‘}}\n\d’))
    s = s[beg: end+2]
    return s
```

Specific information, such as symptoms, duration, causes, risks, medications and mortality, can be extracted from this template. Using the following code, the medications that are typically prescribed for a certain disease are able to be extracted:

```
def find_medication(s):
    s = re.findall(r’medication\s*?=.*<’,s)
    if(len(s)== 0 ):
        s = []
    else:
        s = s[0]
        s = s[:-1]
        s = s[s.find(‘=’)+2:]
        s = s.replace(‘[[‘,’’)
        s = s.replace(‘]]’,’’)
        s = s.split(‘,’)
        s
    return s
```

The World Health Organization (WHO) provides the medical classification codes through the International Statistical Classification of Diseases and Related Health Problems (ICD). Specifically, ICD9 and ICD10 codes can be extracted from the drug information present on Wikipedia. The following code shows how to obtain the medical codes for diseases, signs and symptoms, abnormal findings, complaints, social circumstances, and external causes of injury or diseases.

```
def find_icd10(s):
    s = s.replace(‘|’,’’)
    s = s.replace(‘{{ICD10’,””)
    icd10 = re.findall(‘\w{1}\d{2,6}’,s)
    return icd10
```

CONCLUSION

After all of this data has been extracted it needs a place to be stored. Using a Coma Separated Values (CSV) file is a clean and easy way to store our data. Table 1 is an example on how the UNNI codes can be stored.

Table 1
Stored UNNI Codes

id	name	unii
1912	Ampicillin	7C782967RD
6346	Chloramphenicol	66974FR9Q1
10024	MDMA	KE1SEN21RM
11725	Flunitrazepam	620X0222FQ
14413	Hydrocodone	6YKS4Y3WQ7

Table 2
Stored Diseases, Codes and Meds

id	name	icd9	icd10	medications
4531	Bipolar disorder	['324.0']	['Q273,' 'q20,' 'Q280,' 'q20,' 'Q282,' 'q20']	['Lithium (medication)|Lithium,' ' antipsychotics', ' anticonvulsants']
4581	Bacterial vaginosis	['616.1']	['N76']	['Clindamycin or metronidazole']
4746	Plague (disease)	['020']	['A20']	['Gentamicin and a fluoroquinolone']
5876	Coronary artery disease	['780.0']	['R402,' 'r40']	['Aspirin,' ' beta blockers,' ' Medical use of nitroglycerin|nitroglycerin,' 'statins']

Furthermore, we can store the diseases with its codes and medications using arrays such that it is easier to then access the data when stored as a CSV file (See Table2).

Extrapolating data from Wikipedia can be challenging due to the amount of data it possesses and its inconsistencies. Nonetheless, Wikipedia provides diverse amounts of data that can be used by insurance companies to extract knowledge that otherwise would not be possible. The code provided in the article can be found in a Jupyter Notebook format on GitHub.[2] ■

Alexandru Loan Andrei is a Software Engineer at Reinsurance Group of America (RGA) in Chesterfield, Mo. He can be reached at *Alexandru.Andrei@rgare.com*

ENDNOTES

[1] *https://dumps.wikimedia.org/enwiki/latest/*

[2] *https://github.com/AlexAndrei98/*

The Possible Role of Convolutional Neural Networks in Mortality Risk Prediction

By Holden Duncan

How might a computer tell the difference between the road and a tree while steering a car at 40 miles per hour? A rules system for each possible object that might be encountered could be created; however, such a system only allows for a limited amount of features, and is only as effective as the rules themselves. Increasing the number of rules might make for a more accurate classification, but such an engine would become unmanageable. In addition, attempting to hand-engineer such features limits a model to human intuition of pixel-by-pixel photo recognition.

There are already articles about random forests or linear/logistic regressions, but these methods involve the use of functions and hand-engineered features composed of different categories. As such, these models are generally unable to identify wholly new ideas without specific encoding. However, a model that excels in using the spatial relationship between complex features in data to generate accurate classification could learn to recognize new patterns. A convolutional neural network, or CNN, learns to use concrete low-level features in order to extract identifying abstract ideas. The nodes or "neurons" of the network are organized into different interconnected layers and through training becomes a predictive engine similar to the human brain. While CNNs are most commonly used in image classification, I hope to instead apply them to model mortality risk through visual representations of medical history.

WHAT IS A CONVOLUTIONAL NEURAL NETWORK

Convolutional neural networks are specialized neural networks. Regular neural networks take some fixed-shape input and produce an output. The input propagates through the network via the weighted connections between different layers of nodes, and the transformations which take place in and between nodes produce optimized predictions. Assuming layer A is the layer just before layer B in a given network then:

- Each node in layer A has a weighted connection to every node in layer B.
- Let A_x represent a given node in layer A, similarly for B_y in B, and let $Weight_{xy}$ be the weight of the connection between those two nodes. Then the input from A_x to B_y may be expressed as:

$$Ax_{out} * Weight_{xy}$$

- The net input to node B_y is the sum of the inputs from each node in A to B_y:

$$By_{net} = \sum_{x \text{ in } A} Ax_{out} * Weight_{xy}$$

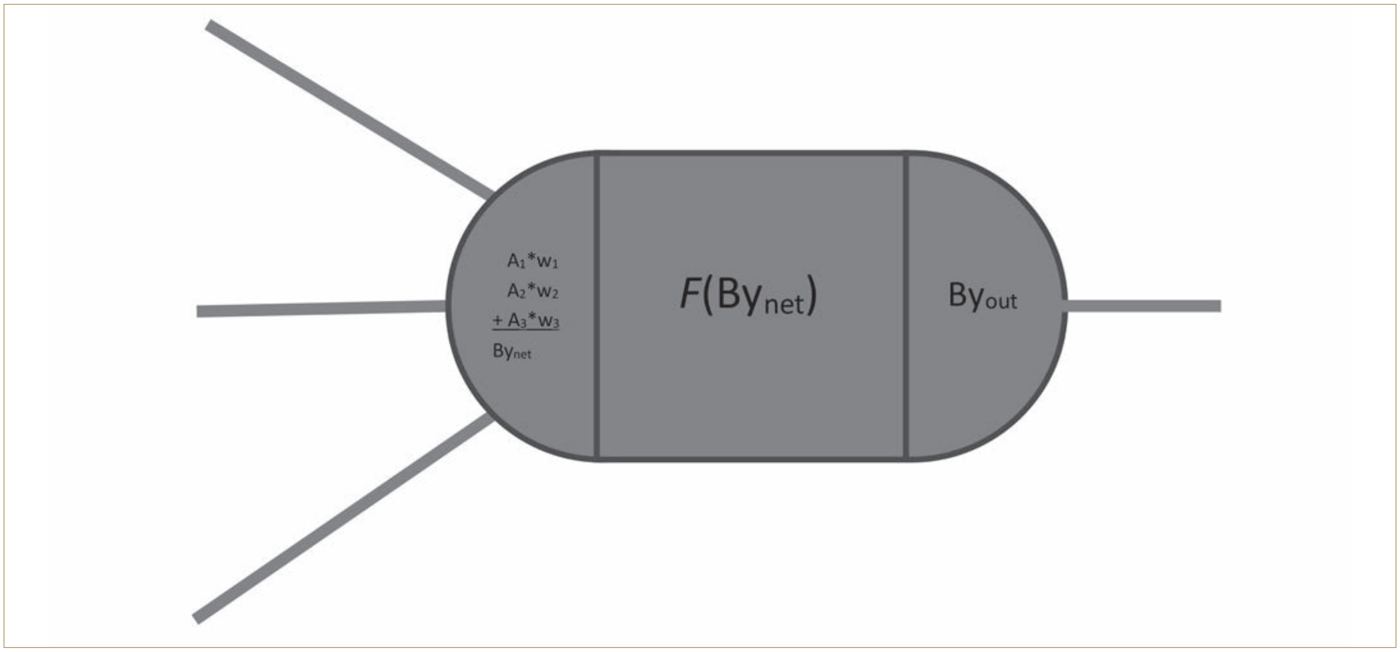
- The output of node By is determined by the result of the activation function, F, applied to the net input:

$$By_{out} = F(By_{net})$$

This process is shown visually in Figure 1 (pg. 21).

Generally, a neural network consists of an input layer and an output layer with any number of hidden layers in between. Each layer may contain any number of nodes. The model makes

Figure 1
Node Output Depends Upon Net Input and Activation Function



accurate predictions using the weight associated with each connection and is able to learn by optimizing these weights. When a fresh network is initialized, these weights are randomly assigned small nonzero values.

Unfortunately, we are unable to hand-wave meaningful weights into existence, but thankfully the network learns through a variant of trial and error and not human intuition. Training occurs through backpropagation, a form of supervised training which compares the network generated output with the expected output by using labeled data, e.g., data labeled "Duck" would have an expected output of 1.0 for "Is Duck" and 0.0 in any other category. In essence, the training data is passed forward through the network and the error is found, and then the network works backwards making adjustments. The Mean Squared Error is commonly used to measure error, thus the total error, Etotal, for the output layer may be expressed as:

$$\sum_{\substack{Node\ in \\ Layer}} \frac{1}{2}(\ target - output)^2$$

An important feature of backpropagation is that changes are applied to each weight individually. The weight of each connection is increased or decreased by the partial derivative of the total error, E_{total}, with respect to the given weight wx. Using the chain rule the error associated with a given weight may be expressed as:

$$\frac{\partial E_{total}}{\partial w_x} = \frac{\partial Etotal}{\partial O_{1_{out}}} * \frac{\partial O_{1_{out}}}{\partial O_{1_{net}}} * \frac{\partial O_{1_{net}}}{\partial w_x}$$

Because E_{total} represents the amount of change needed to reduce the error to zero, the partial derivative with respect to a given weight yields the amount by which that weight must change to minimize the total error. This change is often multiplied by some learning rate to make training more efficient. (These examples assume a learning rate of 1 for simplicity.)

Convolutional neural networks go a step further and use specialized convolution and pooling layers. The output of these layers may be a two-dimensional matrix, or a matrix of matrixes called a tensor. Convolution layers and pooling layers both have kernel and stride dimensions. These variables remain constant within a layer, but may change between layers. The shape of the kernel acts like a spotlight and highlights a limited region of the total input. The region is processed and then the kernel is translated across the input according to the stride.

In convolution layers, the highlighted region is multiplied by a filter. The filter may be matrix or tensor. The sum of the elements in the product represents the similarity between a given region and the filter. The resulting activation map not only shows whether the filter was activated, but also where in the input that filter was activated. There may be any number of filters in a given layer. Multiple filters in a layer will produce a tensor of the various activation maps "stacked" one after another. In early layers these filters detect simple features from concrete input values. In later layers however, filters tend to represent more abstract ideas using the spatial relations between earlier filters. The network operates much like how a human identifies a high-level idea, like a square versus a rectangle, by the low-level information such as the relative positions of each edge.

Pooling layers, on the other hand, serve to only down sample input. A typical method is max pooling, during which the largest value from the input region becomes a single value in a smaller matrix. The result is similar to the input except the location of details are more generalized. Pooling layers are useful because not only do they reduce the dimensions of the matrices within the model, but these layers also help prevent overfitting.

An overfit model begins to simply memorize strict patterns found in training examples. A network trained to detect cars might overfit and expect any detected wheels to be perfectly horizontal. Pooling the layer that detects wheels retains any spatial relationships, but is less sensitive to the exact locations of the features. This generalization makes for a more flexible model.

APPLYING CONVOLUTION NEURAL NETWORKS TO UNDERWRITING DATA

CNNs are not limited to computer vision, however. Because of how these networks use different low-level features to extrapolate complex and abstract relationships, such networks may be used beyond classical images. While the ability to recognize abstract ideas from spatially related features is commonly used for image classification, images are just organized tensors. As such, any tensor or matrix of spatially related data may be used as the input to a CNN.

I have generated visual representations of individuals' prescription histories. The x-axis represents time and the y-axis represents how dangerous the prescription is considered. Darker sections represent a higher number of prescriptions of that severity filled within a given time interval. I have chosen this approach because a convolutional model trained this way may be able to find and use unknown patterns. A note of caution, however: Because most data in columns may be shuffled vertically without the loss of information, tabular data tends not to be a good candidate for a CNN. If one could shuffle the columns of an image and not scramble the picture, then the spatial relationships would be insignificant. The prescription histories, like a picture, have an inherent ordering to the columns. Thus, they can benefit from a CNN analysis.

CONCLUSIONS AND FUTURE DIRECTIONS

The output of a convolution layer with more than one filter is a matrix of matrices. This resulting tensor is passed through the network in order to generate some target output. In order to build, train and test different models, I used Google's TensorFlow library for Python as the backend for the Keras module by François Chollet. The logic for the actual training and creation of the model is based in C and C++ with Google's Python wrapper for interaction. The Keras package is then used to implement TensorFlow as Keras has friendlier syntax and added tools for data manipulation.

Moving forward I plan to use major drug groups, sub groups or even active ingredients in place of severity scoring, thereby possibly capturing new relationships between medications. Theoretically, this network could even diagnose patients through symptom history. In addition, the output may be more than a yes or no answer, but instead a vector predicting the mortality risk of the individual for each coming year. And while I have generated actual images, any two- or three-dimensional input of spatially related data could be used. Such technology is only limited by human creativity and available data. ■

Holden Duncan is a data scientist at RGA Reinsurance Company, in Chesterfield, Mo. He can be reached at *HoldenDDuncan@gmail.com*

The Psychology of Visual Data

By Dorothy Andrews

Accessing the information content of data, large or small, and making it accessible and meaningful is an art form, according to Boehnert.[1] Visualization artists attempt to reduce the abundance and abstract essence of big data to actionable, neutral and objective information. This goal is not easily achieved using data reduction techniques such as line graphs, and bar, bubble and donut charts, all which are subject to axes and scale and manipulation in 2D and 3D. Bergstrom & West[2] point out how graph axes can be used to magnify numerically insignificant heights and lengths of bars in bar charts by excluding the zero origin. By eliminating the zero origin, designers can zoom in on a section of a bar graph to create a visual difference the mind comprehends as more significant than it numerically is. Including zero allows the eye to see the actual heights of all bars and accurately perceive numerical differences. This is less of a concern in line graphs. However, a technique that amplifies insignificant differences in line graphs is changing the scale of graph axes from one where the increments are constant to one where the increments are logarithmic. In common logarithmic scales, each increment along one axis creates a change in the other in multiples of 10, as in measurements of earthquakes. A measure of two is 10 times worse than a measure of one, but a measure of three is 100 times worse than a measure of one and 10 times worse than a measure of two. Mixing such changing magnitudes with non-logarithmic units along the same axis should be avoided to prevent misleading the viewer into seeing a difference that does not materially exist, according to the authors.

Bergstrom & West[3] provide a heuristic to guide the construction of data graphics. "The Principle of Proportional Ink" simply states: "when a shaded region is used to represent a numerical value, the area of that shaded region should be directly proportional to the corresponding value (p.1)." In other words, the area of the ink used to represent a number should be a function of the magnitude of the number and the same function should be used consistently to represent all numbers. The easier the measurement lends itself to a mathematical function, the easier it is to identify and prevent violations of the principle proportional ink.

Filled line charts (line graphs with shading below the line) are akin to bar charts with zero separation between the bars. The authors contend filled line charts should never exclude the zero origin to prevent violation of the principle of proportional ink. Bubble charts can leverage size, color, and horizontal and vertical coordinates to "encode four different attributes (p.5)" of each data element, with some attributes more tractable than others in the design of the chart. Donut bar charts are challenged with conforming to the principle of proportional ink, as well as graphs constructed using changing denominators. The authors point out how easy it is to confound results when percentages rather than absolute numbers are used as the metric in determining the size of graph objects, as illustrated in the graph of causes of death by age groups. While there are more deaths in the 65+ group, the graph suggests (in percentage terms) more 1- to 4-year-olds die of accidental causes. The larger denominator of deaths in the 65+ group is causing this effect. One of these elements could be eliminated from the graph to eliminate the confusion.

Dimensionality is another technique subject to frequent violations of the principle of proportional ink and manipulating perspective to obscure relative sizes of chart objects. Despite a potential ink violation, adding a third dimension can be helpful in visualizing data outliers (data points that are quite distal from the others) that appear to cluster with other data elements in a 2D view, where all the data lie in the plane of a page. A cube view of data adds a height dimension. It becomes easy to perceive a single point that hovers far above a cluster of points as an outlier in 3D than it is in 2D. A purely aerial view can obscure outliers.

Data visualization tools are but one type of big data reduction technique just like the calculation of the mean of a large set of numbers is a data reduction technique. Animation is another and can heighten the experience of visualizing data beyond adding a third dimension (Koblin[4], Lizama[5], Thorp[6], Wright[7]). The "Stop and Frisk" video (Lizama[5]) produced by the Morris Justice Project (MJP) evokes visually as well as emotionally, as it illuminates the geography and frequency of stops and frisks by police in the name of community safety, incurring great costs with little benefit. The video reflects the three elements Wright[7] says are essential to data visualizations: 1) data mining, 2) programming and 3) design. The MPJ collected stop and frisk data on policing in an NYC area neighborhood, programmed the geography of the stops using a flashing light effect, visually simulating a rear-view mirror view of the flashing lights on a police car light bar. Thorp[6] provides several examples of how animating data can reduce its complexity to 2D and 3D to emphasize spatial differences

among data points, allowing for the comprehension of the data from several different "viewing angles." These perspectives are important in differentiating true data clusters from disparate data outliers.

Mack[8] discusses how inattention can distract from perception by interfering with how the brain interprets what the eyes see. The demonstration by Daniel Simon and Christopher Chabris discussed by Mack[8] is a perfect example of this distraction. Viewers are asked to count the number of times a ball is passed among a group of individuals dressed in white or black shorts and shirts. The attention to the task prevents the perception of a man in a gorilla suit weaving among them and exhibiting gorilla-like behaviors. The failure on the part of the observers to see the gorilla results in what the author calls "inattentional blindness (p.1269)." The implication is that attention is necessary in order for perception to occur. Understanding inattentional blindness is important in making sure animated data visualizations do not coerce viewers into seeing less than what is in an image as well as more than what is visually present. The author discusses how the mind will fill in any "blind spots" present in an image. Designers should be cognizant of the potential for the mind to fill in blind spots in such a way that detracts from the intended message of the visualization.

Paying attention to the psychology of visual data can help you use visualizations to inform; and not to mislead. ■

Dorothy L. Andrews, ASA, MAAA, CSPA (Certified Specialist in Predictive Analytics), is the chief behavioral data scientist at Insurance Strategies Consulting LLC. She can be reached at *dandrews@insurance-strat.com.*

ENDNOTES

1 Boehnert, J. (2016). Data Visualisation Does Political Things. Proceedings of DRS 2016, Design Research Society 50th Anniversary Conference. Brighton, UK, 27-30 June 2016.

2 Bergstrom, C., West, Jevin. (2017a). Visualization: Misleading Axes on Graphs. Retrieved from *https://callingbullshit.org/tools/tools_misleading_axes.html*

3 Bergstrom, C., West, Jevin. (2017b). Visualization: The Principle of Proportional Ink. Retrieved from *https://callingbullshit.org/tools/tools_proportional_ink.html*

4 Koblin, A. (n.d.). Flight Patterns. Retrieved from *http://www.aaronkoblin.com/work/flightpatterns/*

5 Lizama, S. (2013, March 19). This Is Our Home: Scars of Stop and Frisk. *Morris Justice: A Public Science Project* [Video File]. Retrieved from *https://www.youtube.com/watch?v=qLWWa2De2b4*

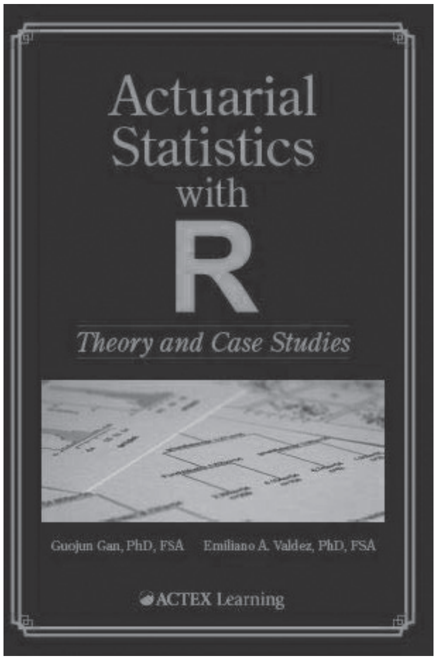
6 Thorp, J. (2013). Numbers that paint the picture [Online video]. National Geographic. Retrieved from *http://library.fora.tv/2013/06/12/Jer_Thorp_Numbers_That_Paint_the_Picture*

7 Wright, L. (2013, December 2). Research Report: Data visualization [Blog]. Retrieved from *http://blog.ocad.ca/wordpress/digf6l01-fw201302-01/files/2013/11/Laura-Wright-Research-Report-final.pdf*

8 Mack, A. (2011). The Image: Seeing More and Seeing Less Than is There. *Social Research*. 78(4). 1263-1274.

Book Review: Actuarial Statistics With R

By Mary Pat Campbell



BOOK DETAILS:

Actuarial Statistics with R: Theory and Case Studies
Guohun Gan, PhD, FSA; Emiliano A. Valdez, PhD, FSA
Publisher: ACTEX Learning
Publication date: 2018

With the hotness of data science and predictive analytics in the actuarial world, many actuaries are looking to get up to speed on the latest approaches. Indeed, I wrote an article in December 2015 about getting started in predictive analytics, recommending various books and online courses.

However, from an actuarial viewpoint, the main problem was that generally one was using non-actuarial examples of data and models to work on. When one is a working actuary, it would be better if one could learn the material and applications specific to actuarial work at the same time.

> With the hotness of data science and predictive analytics in the actuarial world, many actuaries are looking to get up to speed …

This textbook, *Actuarial Statistics with R*, helps to bridge this gap. I received a review copy of this new text from ACTEX Learning, along with a copy of its solutions manual. This does look as if it's developed originally with college students in mind. According to the publisher:

> "The content covers several topics on data analytics that have been prescribed by the International Actuarial Association. In particular, it has been designed to cover the learning objectives for the Statistics for Risk Modeling (SRM) Exam established by the Society of Actuaries. Some materials from this textbook also cover parts of the syllabus for the Modern Actuarial Statistics (MAS-I and MAS-II) Exams of the Casualty Actuarial Society."[1]

While this text is partly actuarial exam related, I'm reviewing this from the point of view of a practitioner wanting to learn some statistical techniques, with a focus on using it for actuarial purposes.

INTRODUCTION TO ACTUARIAL STATISTICAL TECHNIQUES

The book is structured into four large sections: supervised learning, unsupervised learning, time series models and simulation. Within each section, the introduction of the specific technique and its underlying structure or theory is given, followed immediately by one or more case study chapters. For example, the chapter on generalized linear models is followed by three case study chapters: predicting demand for term life insurance, modeling number of auto claims, and modeling the loss severity of auto claims.

I find this case study approach to be very effective. While for college students, these case studies may be introducing unfamiliar insurance concepts, for the practicing actuary, it provides a base of familiarity. In some of these cases, as with regression, many of us had done these sorts of fits before if we've done any modeling, but perhaps we were not so familiar with using R.

Within chapters themselves, there is an alternation between mathematical notation and qualitative explanation of techniques with simple code examples implementing various calculations. The code and data used in each chapter are also downloadable from the publisher's website.[2] It's freely available to the public, though without the context of the book itself, you may get a little confused. There are some comments in the code, but most of the explanation is within the text itself.

The level of mathematical knowledge needed is college-level calculus and linear algebra—there are summation formulas galore, and matrix multiplication makes its appearance. Many of the exercises in the text are more set up for college students or those preparing for the relevant actuarial exams than for the practitioner, but I found the specific symbolic derivations to be important in showing connections between various concepts, such as the connection between the leverage of a point in a linear regression (a measure of how far a point is from other data points, based on the independent variables) and the variance of the residual for a point at that independent variable.

CAVEATS FOR BEGINNERS: DEALING WITH R AND OTHER ISSUES

As the title says, the specific case studies are worked-through in R, and even when introducing the theory, they show R code that calculates the quantities just given via mathematical notation.

The problem starts if you are completely new to R. There is an appendix to provide some introduction, and it is about 50 pages long, so it's not all that bare bones. However, there can be practical issues for someone going to download R[3] and then starting to play with it.

The level of introduction given is still very basic. You are shown foundational data structures and codes, but nothing is particularly complicated. The textbook in terms of technique and code is an introductory level—if you want to start making changes beyond what is given, you will likely need additional resources.

If you are a complete beginner to R, you may wish to look at the R Project's "Introduction to R."[4] From the authors' own listed references, you may wish to look into Albert and Rizzo's *R by Example*[5], which also takes an approach of going through more basic statistical analyses as a way to teach both the language and basic statistics at the same time.

The caveats also pertain to the content in the book in general. This is an introduction to these techniques, providing information on model evaluation and some practical approaches to improving models. There are more advanced texts and materials available once one has learned these basics. It helps to have some experience trying out the basic techniques before trying full-blown approaches.

I recommend those using this text to actually type the code in as one follows along in the text. As mentioned earlier, all the code can be downloaded from the publisher's site, but I think it helps to type in the commands on one's own to think through what is going on. Copying and pasting, or even just running the files as R scripts, means that you may miss some important details.

Finally, a note about some details that I really liked: I liked how the code was presented in the text. As is usual in texts with code, there is a different font from the standard text (specifically a typewriter-looking fixed-width type instead of the more aesthetically pleasing variable-width types). In addition to that, code is set apart in bordered regions, and, most importantly, the lines are numbered.

Given the way the text is set it can be slightly difficult to tell the code from the calculation result (the difference is whether the line starts with > or not), but the lightly-colored line numbers to the left of the code box still helps get one a handle on code blocks that can be as long as 42 lines between code and results.

In addition to a good presentation of the code, there are useful indices in the back—the usual type of index by topic, but also an index of R functions and symbols.

The authors pack quite a bit of information in this relatively short book (374 pages, including indices and excluding front matter), and I think this would be a handy introduction to any working actuary wanting to get started in statistical modeling. ■

Mary Pat Campbell, FSA, MAAA, PRM, is VP, insurance research at Conning in Hartford, Conn. She can be reached at *marypat.campbell@gmail.com.*

ENDNOTE

[1] Gan, Guojun and Valdez, Emiliano A. Actuarial Statistics with R: Theory and Case Studies, page xvii

[2] Link to files are here: *https://www.actexmadriver.com/Assets/ClientDocs/supplement/Supplement_Actuarial%20Statistics%20with%20R.zip* (accessed September 11, 2018). If the link moves, the files will be found via the Product Supplements section of the publisher's website (*https://www.actexmadriver.com/*)

[3] R is downloadable at the R Project for Statistical Computing: *https://www.r-project.org/*

[4] *https://cran.r-project.org/doc/manuals/r-release/R-intro.html Accessed September 11, 2018*

[5] Albert, J. and Rizzo, M. (2001). R by Example. Springer, New York, NY. *https://amzn.to/2Qn4uzZ*

SOCIETY OF
ACTUARIES®

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