Actuary of the Future

May 2017
Issue 40

Learning Data Science From an Actuary's Perspective

By Xiaochuan (Mark) Li

Why do actuaries need to know data science? Insurance is always a data business. Historical experiences are studied extensively by insurance companies to understand risks for pricing and valuation. Actuaries are also well-known for good math skills and meticulous analytical work. Using data is not new to actuaries, who are trained to be good analysts. However, the boom in big data and advanced analytics is driving fundamental change.

We are in an era where the volume of data doubles every three years as information pours in from digital platforms, mobile phones and wireless sensors. Data are also more and more linked to each other. Purchasing history, credit score, driving record and medical information are used to paint the profile picture of an individual, which can be used by companies to provide personalized products and services.

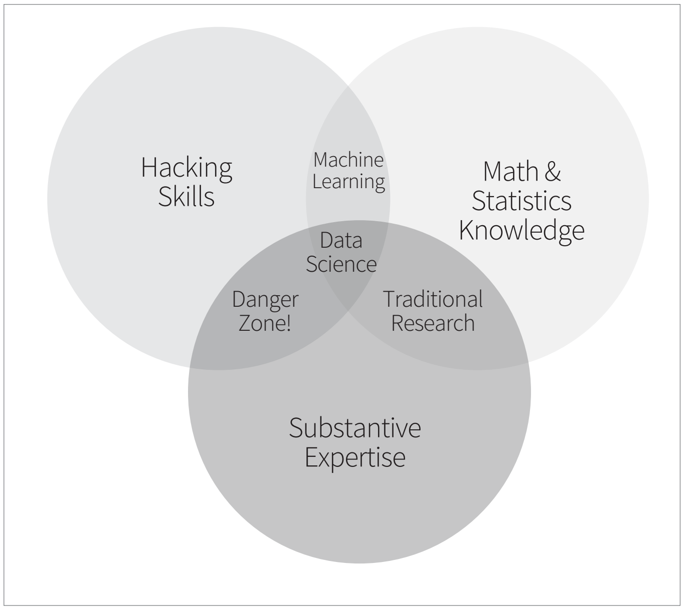
Data produced by customers are already used in other areas of insurance, such as auto insurance, where telematics that monitor driving style help to differentiate safe drivers from aggressive drivers. The development of internet of things (IoT) technology will bring internet connections to nearly every type of consumer device. IoT will have huge implications for the insurance industry. For example, some health care insurers are giving customers free fitness trackers and offering lower premiums or other benefits for meeting daily activity levels.

Insurance companies that are concerned with disruption and innovation are investing heavily in technology. Actuaries need to prepare for the trend as well, as the quantity and types of data will be very different in the future: Variables could be in the thousands or more; data size could be in terabytes or petabytes; data format could be in text, image and video. Unprecedented computing power and more sophisticated algorithms are also available. Actuaries need to understand and to utilize them.

The definition of data science is, however, quite arbitrary. People sometimes use data science, predictive modeling and statistics interchangeably. Some people think of data science as a sexy term for statistics. Wikipedia says it is "an interdisciplinary field about scientific methods, processes and systems to extract knowledge or insights from data in various forms." Although how it is used may not be that important to practitioners, one way to understand the concepts is predictive modeling is a subset of data science, which includes description statistics, optimization, data visualization, data engineering and more. Traditional statistics is one discipline within the several subjects that would have to be used to perform data science.

Figure 1 shows Drew Conway's data science Venn diagram. Since he posted the plot in 2010, people have modified and added more content to it. However, it still covers the major content of the subject. Statistical thinking, programming skills and subject knowledge are still the three main requirements to be a good data scientist.

Figure 1
Data Science



Reprinted by permission of Drew Conway.

MAIN AREAS OF DATA SCIENCE

The following three categories are the main areas of data science, although the divisions may not always be very clear. It is not the complete list, but most activities in data science can fall into these categories.

Statistical Analysis

This refers to traditional statistical methods for prediction, inference and causal modeling of relationships. Causal inference is usually the result of this kind of study. Focus is on understanding the data and relationships. It is usually involved with a hypothesis before proceeding with building a model. For example, linear regression has certain assumptions, like independent, Gaussian distribution, and homoscedasticity of data. The model is not valid if the assumptions are not fulfilled.

Statistical analysis generally can give transparent and explainable results. The downside is that it has conditions and may not achieve the best result.

Machine Learning

Although many machine learning (ML) algorithms are based on statistical theorems, its goal is to make prediction as accurate as possible and allow the data to speak instead of directing it to a specific path with the initial assumption. For example, a decision tree/random forest of the ML algorithm assumes no hypotheses of the data. It learns data, returns the important features, and makes predictions without preconditions.

Sophisticated algorithms can work on large data sets efficiently and can get better results than statistical methods. ML uses data to compute hypotheses that approximate the target. The result may be a black box, and may not be transparent to humans. ML may improve the programs' own understanding, but may not improve human comprehension.

Deep Learning

Deep learning is a branch of machine learning based on neural networks that attempts to model high-level abstractions in data. Neural networks "use a cascade of many layers of nonlinear processing units for feature extraction and transformation. Each successive layer uses the output from the previous layer as input."[1] The nonlinear processing units are usually called neurons, inspired by biological neurons, which can transmit an electrical signal from one end to the other from dendrites along the axons to the terminals if the signal power reaches a threshold.

Deep learning allows computers to build complex concepts from simpler concepts. For example, in image recognition, some neurons represent shapes; some neurons represent colors; and some neurons represent edges. The whole network can combine these concepts represented by individual neurons to represent the concept of an image. In the past few years deep learning has achieved huge successes in computer vision, natural language processing and artificial intelligence. A Japanese life insurance company has begun to use IBM Watson, which is based on deep learning, to process claims.

BIG DATA TECHNOLOGY

Actuaries are good with Structured Query Language (SQL) to manipulate data in relational databases. However, this type of database has disadvantages: It is mostly for structured data; and it can be very slow to process large amounts of data.

Apache Hadoop is an open-source software framework used for distributed storage and processing of very large data sets and processing using the MapReduce programming mode. It can store structured and unstructured data. It is more scalable and much faster than a traditional database. Apache Spark is another new technology that enables distributed in-memory computing. It can work on the top of a database or Hadoop to provide very fast computing performance. These technologies have already been adopted by many insurance companies.

LEARNING VENUE

Traditional actuarial education focuses on actuarial mathematics and insurance risk. Students are often trained to learn how to use Excel spreadsheets or SAS. Although they are also used in data science, data science additionally requires more advanced levels of linear algebra, statistical theorems and computer programming languages. R and Python are two of the most common programming languages used in data science.

If you are still in college, you may sign up for courses on these subjects offered in computer science and statistics departments. If you don't have access to these classes at school, there are tons of free or affordable resources on the internet. Major massive open online course (MOOC) websites such as Coursera, edX, and Udacity are offering courses in statistics, machine learning and programming languages. Most courses are at the introductory or intermediate levels. They are often taught by professors from the best universities in the world. For example, many people entered the data science field by taking Andrew Ng's machine learning class on Coursera. Khan Academy is also offering classes at a more introductory level to help students learn fundamental theorems in statistics and linear algebra.

Universities also have their online programs, and many of them are free. For example, Stanford Online offers advanced levels of deep learning classes on computer vision and natural language processing.

YouTube has many channels and videos that focus on the data sciences. Conferences on Python and R, like PyCon and Rconference, are posting presentations there. It is good to learn from experts in different industries on how to use these tools.

Data science is a fast-developing field. New methods, algorithms and tools are coming out. Social media is the best place to stay current with new developments. To get updates, you can subscribe to relevant blogs and tweets. R-bloggers, Data Science Central and KDnuggets are popular websites that have valuable materials.

Quora is a question-and-answer (Q&A) website. It is also a good resource to ask a question and learn from others' answers. People tend to answer complex questions in layman's terms.

PLAYING FIELD

Data science is a very hands-on subject, and the quickest way to learn it is by practicing it. Kaggle is a website dedicated to data analytics competitions. It draws many data scientists from all around the world to compete to produce the best models. It provides problems, data and prize money. Most of the data are cleaned real-world data. You can get a flavor of different types of data in different industries. It also promotes collaboration and learning by encouraging people to share models and programs and discussions, which is probably the best resource for beginners. Companies also find participation attractive. Many big companies, including insurance companies, are sponsoring the competitions and even treat them as a recruiting opportunity to hire the brightest data scientists.

CONCLUSION

For actuaries, learning data science does not necessarily require becoming a data scientist to implement algorithms from scratch (great if you can!). There are tools and frameworks of data science available that provide easy-to-use interfaces to sophisticated algorithms without reinventing the wheel. This knowledge is more likely to help actuaries have a better assessment of problems and solutions. Nowadays lines between different subjects are becoming blurry. It is always good to be open-minded and stay curious.

Good data scientists need subject expertise. Actuaries are in a unique position to embrace data science, rather than consider it as a field that is in competition with the actuarial profession. Data science can push the profession and the industry forward to face new technology and a changing world. It will enable actuaries to grasp more tools to analyze data, and it will definitely boost career potential and opportunities. ■

Xiaochuan (Mark) Li, ASA, is a data scientist at RGA Reinsurance Company, in Chesterfield, Mo. He can be reached at *xli@rgare.com.*

ENDNOTE

[1] *https://en.wikipedia.org/wiki/Deep_learning.*