Learning To Be a 'Rocket Scientist of Statistics'

Date : September 26, 2017

Aneesh Shinkre is a senior at Allen D. Nease High School in St. Augustine, Florida. This summer, Shinkre spent six weeks as a business analytics intern with Latize, a start-up data management company that uses a platform known as Ulysses, a software that helps customers “realize big data’s huge potential” by combining internal and external data sets into “an intelligent web of data.” During his internship, Shinkre dove into that data from a programming perspective and began to see how it can be used to solve big problems, like financial fraud. In this personal essay – part of the KWHS Summer Essay Series — Shinkre writes about credit cards, coding, and his new appreciation for data science.

Picture this. It’s 3:30 a.m. and an unassuming consumer is fast asleep, snoring hard. Little does he know that within minutes, 10 different credit cards in his name, each with no spending limit, will be shipped to a network of ruthless fraudsters halfway across the globe. They will then use the credit cards to spend, spend, spend. What’s about to happen to this slumbering consumer happens every single day, in the millions – in fact it’s happening in the hundreds as you’re reading this article. It’s called credit card fraud and identity theft.

The good guys of the industry are fighting back against these scenarios even stronger than ever, equipped with the defense of artificial intelligence and machine learning.

‘Rocket Scientists of Statistics’

Every time someone swipes a credit card, writes a check, withdraws money, or makes any form of a transaction, it’s recorded in a database of billions upon billions of transactions. It’s just sitting there, filled with 0’s and 1’s, yes’s and no’s, customer addresses, and random dollar amounts. Mining, cleaning and ultimately analyzing this data could prove to be a powerful tool that has the potential to end many types of financial fraud.

Data scientists, whom I like to think of as the rocket scientists of statistics, are among the most sought-after job titles these days. The Bureau of Labor Statistics estimates that data scientists, data engineers and data developers will fill 700,000 new jobs by 2020, with the average position pulling in an attractive six-figure check a year. As our society goes more digital, the sheer volume of information and activity being recorded is only going to skyrocket. Someone has to analyze and work with all that data.

Personal finance generates many of these numbers. Increasingly, startups, businesses and programmers (like myself) across the world are trying to detect patterns in each of these data points to determine if a transaction on a credit card is clean or fraudulent. For instance, let’s say we have a sample data profile of a married American who makes $45,000 a year, owns $150,000 in assets, has three credit cards with a consistent credit score, and all of a sudden a charge occurs at 4:22 a.m. for a purchase costing 8,000 euros. As soon as this data point gets recorded in the credit card-issuing bank’s system and is linked to the client’s personal information and spending history (such as last purchase or level of account activity), the computer immediately raises a red flag. Why would someone with that level of income buy something at that unusual hour for that much money, all in a foreign currency?

Once the machine detects the suspicious data, it identifies it as fraudulent. This is where the intensive programming comes in. We’ve got to program the machine to make sure it can tell right from wrong, or in this case a typical transaction from criminal activity. This involves writing code to say “mark transaction X,Y, and Z as clean, and mark A,B,C as fraudulent.” Remember, a computer can’t think or reason for itself — it just does what you tell it to do. But boy, does it do it fast, with billions of calculations every second! So, once you’ve taught this machine right from wrong through various examples (known as a belief system), it's time you put it to the test in the real world. We're working on generating thousands of these fake data profiles, as previously mentioned, through a programming language known as R. As we feed more and more of these data points to the machine, it gets increasingly powerful and observant about real-world activity. As the platform gets more exposed to the real world and as we tell it right from wrong, it may possibly get smarter than us one day — automatically!

Getting into the Numbers

As a newbie to the field of data science during my internship with Latize this July and August, I was faced with a massive learning curve. I was intimidated and also exhilarated as my high school brain saw statistics through an entirely new and AI-supported lens. I was able to work on my business analytics internship from home, so a lot of my time was spent researching key objectives, learning new programming concepts, and familiarizing myself with the analytics jargon in preparation for conference calls to discuss how we'd facilitate the learning of the machine so it could tell right from wrong.

My main job was centered on creating fake data profiles with names, addresses and other information to test. The company had a platform up and running, and they wanted to test it based on sample data. This is mainly where I used R and Excel to randomly generate data, purposefully incorporating true negatives, false positives and any data points that might come in between. If the computer platform responded correctly to any random, hypothetical person I created — including their name, address, ATM ID, transaction ID, credit card number, and so on — then it would get validated by the system. If it misidentified a fraudulent transaction, however, we then made sure that the machine was made aware of its mistake so that it wouldn't happen in the future. This process is similar to how self-driving cars and email spam detectors are tested before they operate in the real world.

As my understanding grew exponentially, the connections I was making seemed to be subconscious and spontaneous. Still, the work was new and challenging. When you're unfamiliar with the general environment of how things work, sometimes your best bet is to acutely observe what's going on, how problems are solved, and try and mimic that in the best and most accurate way possible. In the world of data science, and in many other practices, the best way to learn is by trial-and-error. Although the error can be disappointing, just remember that you've not failed, but you've found a thousand ways that don't work, according to the wise words of inventor Thomas Edison. The theory is always going to be available at your convenience, but the application and experimentation of the theory are what move society forward. I learned so much over my six-week internship, and I spent the rest of my summer dreaming in multidimensional vectors, statistical distributions and computer code.

Big Business Ethics

While I discovered the power of big data this summer, I also saw the potential for things to go very wrong. As much as we wish to do a favor with our extensive business analytics, we must proceed with caution. For example, my team had to ensure that we were doing society a favor with our data analysis, rather than taking advantage of innocent customers. We had access to millions of data files and transactions, as well as a machine with a massive, complicated network of ideas. If the algorithms didn't function the way we intended, we could have been mismatching fraudsters with everyday people. Our team had to be led by business ethics. I realized that it's always a great idea to zoom out, see the big picture, and really ask questions like "What's the end goal? What problem am I trying to solve here? What happens if I link certain data files to other attributes?"

One of the greatest skills I picked up during the course of my internship was the need for a start-up business to identify and focus on society's problems – that pain point, like credit card fraud. Startups must provide something that other bigger companies don't by tackling or even creating a specific niche in the marketplace.

Consider this: we have 1.2 trillion terabytes of data hidden and lurking around the dark avenues and corners of the internet. Each one of those trillions of terabytes tells society a story: what kind of customers like X brand of potato chips, what socioeconomic demographic is most likely to binge-watch Netflix shows — anything the human brain can imagine is already sitting in databases, eagerly waiting to be mined, cleansed and analyzed.

You don't need the brain of Einstein or a master's in computer science to start working with that data. All you need is some elementary school math and statistics, paired with a passionate desire to explore the unknown. With some basic comprehension in programs like R and Python, you're heading in the right data direction. The numbers are, and will always be out there to analyze and decode. Who's up for the adrenaline-filled journey of solving the puzzles and connecting the dots?