August 23, 2016

Don't Be a Big Data Snooper

Alex Woodie

One of the biggest challenges that data scientists face is separating true predictors from false ones. When an airtight causal model can't be created, data scientists often look to a secondary class of models based on correlations to accurately predict outcome. However, when using these models, great care must be taken to avoid falling victim to the data snooping bias.

Data snooping is essentially the practice of finding patterns in data that don't actually reflect the real world. Data scientists may know it by other names, like overfitting the curve or confusing the noise for the signal. The simple definition makes it sound like data snooping would be fairly easy to avoid. However, because of the way the human brain works and how it's wired to spot connections in seemingly disparate pieces of data and events, it's one of the most difficult biases to eliminate.

Data scientists are particularly prone to data snooping bias when they're doing freeform exploratory data analysis, as opposed to attempting to prove or disprove a hypothesis before digging into the data. Traditionally, the best way to eliminate the data snooping bias is to institute strict controls in their experiments before they begin. Chasing interesting results once the experiment has started is a good way to fall victim to the snoops.

Over the years, data snooping has been one of the toughest biases to correct for in the world of applied statistics. In particular, data scientists and statisticians who work in the financial field are more prone to data snooping than in other industries, argues MIT professor Andrew Lo.

In his 1994 paper "Data-Snooping Biases in Financial Analysis," Lo wrote:

"Given enough time, enough attempts, and enough imagination, almost any pattern can be teased out of any data set. In some cases, these spurious patterns are statistically small, almost unnoticeable in isolation. But because small effects in financial calculations can often lead to very large differences in investment performance, data-snooping biases can be surprisingly substantial."

In our current big data age, where data scientist and analysts are building all sorts of models to explain and predict how the world around us works, it's safe to say that data snooping is as big a problem as ever.

Fighting Data Snoopers

One data scientist at the forefront of eliminating data snooping bias is Ryan Sullivan, the CEO and founder of a San Diego, California-based data analytics firm called Intensity. In the 1990s, Sullivan and his UCSD professor, Allan Timmermann, published details of new data snooping bias techniques that could reliably separate the models with solid statistical foundations from those built on sand.

"One of the key issues with big data is not whether we have enough data, but identifying that which is predictive versus that which isn't," Sullivan says. "Part of my claim to fame is that I, along with some of my colleagues, pioneered the application of statistical techniques that allowed one to identify truly predictive factors and truly predictive models from those that aren't. There are some that look predictive, even though there's no underlying basis for it."

In one famous example, an economist searched for a reliable proxy for predicting the S&P 500 index. "He determined that the best one was butter production in Bangladesh," Sullivan says. "He did this all tongue in cheek to demonstrate that, if you look hard enough, you're going to find some factor that looks predictive but obviously has no relationship. Clearly, butter production in Bangladesh is not going to help us predict the S&P 500."

Lo had another good example of data snooping in his 1994 paper that involved interesting mathematical principles behind prime numbers. There's a class of numbers that mimic some of the odd behavior of primes that are called "Carmichael" numbers. It turns out there are only seven Carmichael numbers between one and 10,000, and that if one were to pick stocks based on those equities that had Carmichael numbers embedded in their stock identification numbers, one would have done abnormally well.

Causal Models, Supernatural Connections

Of course, there's no real-world basis for why those numbers did well. It's just dumb luck. But that won't stop unscrupulous salesmen from playing on people's belief in supernatural powers and extraordinary metaphysical connections. The human desire to have inside knowledge is powerful, and plenty of snake oil has been sold through the ages because of it.

Data snooping can also affect those analyst and data scientists who are entirely above-board with their intentions, but perhaps just not as scrupulous in their methods as they should be. Sullivan developed his anti-data snooping techniques to weed out the bogus models so the truly predictive ones can shine.

"When we're dealing with big data, we have lots of data that we're trying to evaluate, and we can put that through a gazillion types of models," Sullivan tells Datanami. "But we need to have some way to correct for the biases that

naturally result [when we're] finding something that looks predictive, to be able to determine whether it is or not."

The gold standard among both economists and other analytic professionals is a causal model that incorporates the fundamental factors influencing a given system. However, in the modern world, causal models can be difficult to build, Sullivan says.

"A causal model can deliver solid performance if one can truly identify the causal factors and measure them well," he says. "But there's a lot of 'ifs' in doing so. That's why a predictive framework can be much more efficient and effective, because there can be a predictive factor that's truly a reliable predictor without it being a causal factor. Often times we can't measure or see a causal factor."

There are other advantages to using predictive models compared to causal models, including the fact that predictive models are more flexible and enable users to do "what if" type of scenario building. Those are tough to do in strictly causal models, he says.

"A good causal model is really difficult to develop and typically is not flexible," Sullivan says. "Because we have a predictive model we can better evaluate the future uncertainty. We can measure and quantify the range of potential future outcomes in much more accurate fashion."

"The gold standard among both economists and other analytic professionals is a causal model that incorporates the fundamental factors influencing a given system."

In a complex world that has a lot of factors at play, a good predictive model can sometimes give the best answer—provided the biases are accounted for, of course.

"The issue is those relationships are not simple correlations, but can be highly dimensional," Sullivan says. "It can be the intersection of multiple factors that give rise to something that's predictive. And that's where the sophisticated techniques and algorithms allow us to identify those relationships that would otherwise go unknown."

Sullivan's firm uses ensemble modeling techniques to ensure that the models get better with time. "We're continuously taking feedback from the errors that arise in our models, because of course all of our models are just that: they're models. They're predictions," he says. "They are not perfect. So we identify the errors and measure those and thus are able to continually improve the forecast as we go forward."

At Intensity, Sullivan and his team—which includes his former professor, Timmermann, one of the world's foremost authorities in economic modeling—have created what he claims to be some of the most powerful and accurate economic models available. The predictive models use a variety of public data sources as inputs, including industrial production, employment rates, interest rates, and GDP. The models are updated continually on the Microsoft and Amazon clouds.

According to Sullivan, three factors make the modeling framework possible: nearly unlimited computing horsepower, large amounts of data, and an experienced team of analytic professionals. "It's not easy. It really is not easy at all," he says. "That's where the challenge of assembling a really good team and merging that with the breadth of data, with the computational horsepower–all of those things come into play."

Predictably Unpredictable

One area where the models must be revisited continually involves one of the least predictable areas: consumer buying behavior. The challenge is, it's almost impossible to directly measure consumer sentiment.

Predictive frameworks can be useful for understanding phenomenon that can be tough to directly measure, like consumer sentiment.

"But we can measure the factors that directly influence that buying behavior," Sullivan says. "And as a result, we can bypass the measurement of the consumers themselves, and measure that which directly impacts their behavior, and as a result impacts and predicts the performance of a company."

Building such a predictive model is no easy task, which is why data scientists make the big bucks and why the field of big data analytics is rife with stories of failed projects. Trying to assign too much scientific certainty to a process that is "predictably unpredictable," like consumer purchasing decisions, can drive you mad. That may be when you need to hire what Calabrio senior vice president Matt Matsui recently dubbed a "data whisperer."

"People get hung up on math, algorithms, models, numbers, and data, and the truth is all those are just proxies for trying to predict and understand human behavior," Matsui told Datanami earlier this year. "That's the part that gets lost in this so often–that all those numbers are really in service to try to predict something that's predictably unpredictable."

For those who are looking to generate the best possible predictive model—which is hopefully most Datanami readers–then being aware of, and trying to eliminate, the data snooping bias is definitely a worthwhile goal. The challenge is that data snooping is involved in how we see different pieces of data being connected, which at the end of the day is a fundamental urge of the human race.

"When one is doing data mining, one is hunting for relationships," he says. "Unfortunately that sometimes can go down into the world of data snooping, which then gives us things like butter production in Bangladesh, that ultimately are useless. It's a matter of being able to filter those out, which is where the advanced techniques come in. Certainly it's an area of focus for many, many folks who are dealing with big data."