Module 11: Practical Application 2

Video Transcripts

Video 11.1: Practical Application 2

Welcome to your second practical application module. Just like last time, there's no new content for this module, but there are career components before, and a mentoring component after this module.

The career component complements this time to focus on LinkedIn and networking, as well as navigating the job search. The mentoring component will provide guidance on how to tailor the concepts covered in the career components to a job in ML/AI.

For your practical application this module, you will explore a dataset from Kaggle (an online community of data scientists and machine learning practitioners) that contains information on three million used cars. Your goal is to understand what factors make a car more or less expensive. As a result of your analysis, you should provide clear recommendations to your client, a used car dealership, as to what consumers value in a used car.

Video 11.2: Prediction Problems

Today, I want to think about when and where we can use the powerful predictive tools we're learning in this class. Because our metrics for model selection and development focus on out-of-sample fit and other measures of predictive performance, it's natural to think about prediction as a key skill in being an effective data scientist or model developer. To some extent, this is true. But we should also be aware of the limits of prediction and how to assess when we need other tools for data analytics. The main question you should ask yourself in commencing on a data science project is whether the question you're trying to answer is in the world as we know it, or if it is a what-if question.

Machine learning is very good at developing models and insights in settings that are sufficiently similar to that in which the model is trained. For example, if we want to know what kind of ad someone will click on in an email platform, and we have a lot of data on that person and people like them in the same setting, we know prediction can do quite well. On the other hand, when we're doing something new, say offering a new product, bringing a new drug to market, or changing a user experience for a website completely, using past behavior to predict what might happen is not very effective. This is an important limitation to machine learning and why it is only one part of the toolkit a data scientist needs. Instead, we want to use causal inference methods, such as randomized control trials or AB testing.

I want to introduce this issue by thinking about an important opportunity for applications of machine learning to improve people's health. In many countries, but in the U.S. in particular, a lot of health care is consumed by a very small number of very sick people. It also turns out that many of these people are not just sick. They have many other challenges in their lives — such as low incomes, unstable housing — that make taking care of themselves difficult. This creates an opportunity for policymakers and entrepreneurs to try to better allocate costly scarce resources earlier to these individuals to improve their health and lower the cost of care provided to them. This model, often called the Hotspotter model, was pioneered by Dr. Jeff Brenner in Camden, New Jersey, and has since been at the core of numerous efforts by public health care systems and startups such as Cityblock Health.

While the approaches differ, all of them have some form of effort to identify high utilizers and enroll them in programs that help with many aspects of their lives as well as helping them with health itself. Take a moment to think about what prediction problem you might face in this context. What kind of data and tools would you want to build a model to predict who should be in the program? How might they benefit? Hopefully, you can see that data assets in machine learning are a potentially important piece of the model. It's not enough though, to simply use the data that we have to make predictions and evaluate whether this is a good policy or business. Why? Because we need to know what would happen if someone were enrolled in the program. It is not enough to simply predict a high cost and see whether it is reduced.

To see why, we can look at a very nice study done by Amy Finkelstein and colleagues in the New England Journal of Medicine. They actually put the Hotspotter model to the test. They enrolled patients in the hospital with very high spending into the trial. These patients were then randomized into either care as usual, or the high touch interventions of the Hotspotter model. They show that the cost of care for those patients who were put in the Hotspotter model drops dramatically following enrollment in the program. Within six months, those patients are admitted to the hospital only about half the time. Seems promising, right? Unfortunately, when they add a control group, people who were also sick but did not get the program, things look far less rosy. Those without any intervention only saw their cost of care reduced by exactly the same amount, as well as the same reductions in hospital admissions, and days in the hospital. The important issue they show is that when you identify risk at a point in time, people naturally revert back, either because they got better on their own, or treatments were effective. This is important news if you're running an insurance company or a health system, or are an investor in this area.

But all is not lost for the data scientist. There's still a very important prediction problem that this study and looking at the data raises. If the model is going to work, we need not just predict illness or high cost, we want to predict which patients can actually benefit. This means we need to identify consistently high-cost enrollees. This is a different problem, but one that we have the tools for, and can get the data for. It remains to be seen whether a better algorithm to identify patients can be effective. But in data following a small sample of 2,500 high health care utilizers for two years, only about 350 were consistently high cost. This is promising that we can make an effective model work. It also underscores the need to carefully evaluate programs and business models and link model development closely to outcomes that can be verified and provide value.