How can machine learning methods help take the 'con' out of 'econometrics'?

In this EconTalk episode with guest speaker Susan Athey, the Stanford professor described the danger of researchers misusing machine learning to create self-fulfilling prophesies. She then offered several solutions to combat this classic perilous tendency in econometrics.

Correlation is not equal to causation. This is the one of the main points Professor Athey wanted to come across. Raising minimum wage is a classic example. Many policy makers want to know the effect of minimum wage on employment. If one just conducts a simple regression between a city's minimum wage changes in the past five years and the city's corresponding employment level changes, chances are, he or she would find some kind of correlation. However, this correlation cannot prove that minimum wage has causal effect on employment, because within that five-year period, a lot of other conditions of the city probably have also changed, e.g. GDP growth, working population age. Therefore, it is necessary to control all other variables except the target variable while performing regression, which is the idea of "difference in difference" brought up by Athey. For example, if one wants to research on raising minimum wage's effect in Seattle since last year, one should collect a number of cities that underwent similar economic situations like Seattle in the past year, except that these cities did not raise minimum wage. Then, by comparing the employment rate change in Seattle and that in controlling cities, one can make the conclusion if causal effect exist or not. The reason "difference of difference" is important is that if a researcher tries hard enough, let's say 1000 regressions, he or she will always find a regression to support the conclusion he or she wants. It is crucial to show, in this case by "synthetic control method", that the conclusion is not only related to a particular dataset.

Machine learning can either exacerbate or ameliorate the creation of self-fulfilling prophesies in econometrics, depending on how researchers choose to use it. Athey introduced several steps researchers should adopt to mitigate biased in machine learning. First, Athey argued that, instead of making research assistants handpick 50 models to try out and select the model whose result best supports the research hypothesis, researchers should let the machine run all models with no pre-established bias, and then pick the best model. The most important part is the second step: always split the data, using one half to figure out the model and the other half to fit the model. The reason is that you do not want your model to be only related to this particular training data, i.e. you do not want the explanatory variable to have a high standard error. That is to say, the predictive model(the first half of the data) shouldn't produce causal conclusions. To prove causal effect, additional pseudo-test data is indeed needed.

Xiaoli Jin