CAN MACHINE LEARNING AID IN CARTEL DETECTION?

BY ROSA M. ABRANTES-METZ & ALBERT D. METZ[1]

 

1 Rosa M. Abrantes-Metz is a Director in the Antitrust, Financial Regulation and Securities Practices at Global Economics Group, and an Adjunct Associate Professor of Economics at New York University's Stern School of Business; RAbrantes-Metz@GlobalEconomicsGroup.com. Albert D. Metz is a Managing Director at Moody's Investor Services; Albert.Metz@moodys.com. The views expressed in this article are our own independent views and do not represent the views of the organizations with which we are affiliated or our clients.

I. INTRODUCTION

Recent research has focused on complex antitrust issues stemming from corporate uses of "Big Data" and "Machine Learning" pricing algorithms. For instance, could the pricing algorithms of two different companies ever be said to be colluding with each other? In this short note we want to explore the other side of the coin and ask whether Big Data and Machine Learning could be used in the detection (and therefore in the defense) of cartels or other collusive, anti-competitive practices, and what, if anything, would be the role of the economist in such applications.

However one wants to label the application of sophisticated pattern-matching algorithms to large data sets – "data mining," Big Data, "artificial intelligence," "machine learning" – it is often considered to be a field of expertise separate and distinct from traditional economics or even econometrics. We will not spend time developing a taxonomy over these different concepts (that itself being an interesting and nuanced exercise) but will simply refer to these collective practices as "machine learning," a field (or maybe set of fields) often considered the domain of data scientists or computer scientists much more so than economists. Does this field have a home in cartel detection, and can (or should) the economist be excluded?

II. WHAT IS MACHINE LEARNING, AND HOW DOES IT DIFFER FROM ECONOMETRICS?

A useful if not rigorous definition of machine learning is that it is an application of minimal-structure pattern-matching algorithms to (i) infer a classification rule from a training data set and (ii) make useful predictions on new data.[2] Of course one need only dip one's toes in the water to see that this grossly over-simplifies the field, but it is fair to say that the primary goal of machine learning is to predict, or really classify. Is this an image of a human face or not? Is it an image of John Smith or not? Based on other users, what is the most helpful item to return if one enters "machine learning" in a search engine? From past experience, should we classify the demand for ride sharing to be "high" or "low" tomorrow afternoon (note that while there is a football game scheduled, severe thunderstorms are expected)?

These algorithms can be very powerful predictors. That is their purpose. It has been established both in the rich theory of machine learning and in the practice of our daily lives, where we are confronted with these applications constantly and, at this point, seamlessly. Of course they are not perfect, but very often they are extremely useful.[3]

If we compare this with traditional econometrics, which may be defined as the application of statistical methods to economic problems, we immediately see important similarities, but also important differences. Econometric methods have long been used for purposes of predicting outcomes. The numerical techniques associated with "linear regression" always produce what is known as the "best linear predictor" even when some of the assumptions of true regression analysis do not hold.

We are therefore not surprised to find these same numerical techniques in textbooks on machine learning. Whatever else we want to say about Big Data and its novelties, any data scientist who wants to restrict her hypothesis class to linear functions will end up doing the same numerical procedure that an econometrician would. What is typically not found in those textbooks is the discussion of the statistical and distributional theory of regression which, after all, is largely beside the point of the machine learner.

Yet this statistical theory, and not the value of prediction, is really the emphasis of econometrics. Under what conditions can we test a hypothesis about how an outcome Y relates to an explanatory variable X? That knowing X would allow us to form useful predictions of Y is sometimes seen as a "nice-to-have." Econometricians sometimes take pains to argue that measures of fit, like the well-known R2 statistic, are of secondary or even tertiary importance. An econometrician may declare success in his research even with a very low R2, because – quite rightly – a high fit isn't necessary for the theory to be correct. But for those interested in prediction, it is likely of primary importance. A famous aphorism among econometricians holds that if you want to predict the number of left shoes sold, get data on the number of right shoes sold. You will have a perfect prediction, but you will have explained nothing, you will have nothing to test, and in fact, the conditions necessary for regression are altogether violated!

2 Some might argue that to truly be a *learning* algorithm, the system must be able to update itself given new data. We consider it beyond the scope of this note to distinguish truly adaptive from static algorithms. The essential point for us is that the algorithm may be described with very little structure, and hence generally requires both large amounts of data and substantial computing power to train.

3 Indeed, the theory of learning often speaks in terms of "bounding" the "probability" that an algorithm will be "probably approximately correct."

III. COULD MACHINE LEARNING ENHANCE AND SUPPLEMENT THE EMPIRICAL DETECTION OF COLLUSION?

The general task of "collusion detection" (by which we mean any form of anti-competitive behavior detection) would seem to be a problem of prediction or classification to which machine learning would be well-suited. One could argue that all we seek to do is determine whether a certain arrangement should be classified as "collusive" or "not," just like classifying whether a given set of pixels should be classified as a "human face" or "not."

The answer is yes…but. A machine learning algorithm requires a training data set: to train a machine to detect collusion you have to show it what collusion looks like and also what non-collusion looks like. This training set must be of sufficient size (to guarantee, at least probabilistically, satisfactorily low error) and contain correct classifications of the outcomes as "collusive" or not. Does such a data set exist today – a data set with a sufficient number of cases of both collusion and not-collusion, with the necessary data on price, cost, and drivers of supply and demand – or will we have to wait for it? And how precise is the classification of "collusion?" Presumably it must be correct as to when it started and when it stopped. That is sometimes well-defined and known, and sometimes a bit murkier. What if the collusion starts with two, and then adds more members over time – when should we teach the machine to classify as the beginning of the collusion? What if one member is cheating slightly at some point? Should the classification of "collusion" be a continuous measure (the "strength" of the collusion) rather than a binary yes/no?

Finally, how useful would a classifier of "collusion" versus "not" truly be? The legal requirement is an identification of explicit collusion. Yet it is well understood that empirically, tacit collusion can be virtually indistinguishable from explicit collusion. What then would we train the machine on? If we ask it to classify "explicit collusion" from "not," that might very well be hopeless, since from the machine's point of view there would be data that was essentially identical yet classified as "not." The missing factor would be data on "did the parties explicitly agree to collude." That information, obviously, would allow the machine to separate cases of explicit from tacit collusion. But just as obviously, that is the only piece of information that would be required, and such a prediction rule – "classify the case as explicit collusion if there is an explicit agreement to collude" – is utterly useless in practice.

If the idea of having a universal machine algorithm which could detect "collusion" from "not," to say nothing of detecting explicit collusion from not, seems a bit out-of-reach today (and we stress that we don't know, even if our skepticism is showing a bit), other related problems may be attainable instead. For example, it might be possible to train a pricing algorithm on a set of data from one period and establish that it is not predictive of prices from a different period. If that is the case, can we say that we have detected collusion? The argument will come down to whether there was a structural break (e.g. the establishment of a price-fixing cartel), or whether it is just a bad pricing algorithm. As another example, one could perhaps train an algorithm to identify if prices become somewhat unresponsive to costs, or if prices become more tightly clustered across firms. In short, there may be classification problems circumstantially related to collusion which might be more susceptible to machine learning.

IV. THE ROLE OF THE ECONOMIST

To resort to a tired cliché, to the economist the use of machine learning and Big Data to detect a cartel is more evolutionary than revolutionary. Economists and regulatory agencies have been using data to empirically screen for cartels for years. As data sets have grown larger and computing power has improved, the popularity of non-parametric or "unstructured" techniques has increased. When the data crosses the line from large to Big, and when the methods become "machine learning," we will leave for others to sort out. But clearly there is a continuum, and economists have already been moving toward less structured, more data-intensive methods as they are able.

Almost two decades ago, before "Big Data" was a phrase, one of us worked with the compliance department of a services company to identify which managers were colluding to boost each other's performance evaluations. In this company, each manager was asked to rate the others, and it was suspected that a group had agreed to boost the scores they assigned within and depress the scores they assigned without. Given the anonymized data on how the managers had classified each other and other relevant characteristics such as their location, their area of practice, and others, and data for several years, we used a clustering algorithm to run over all possible combinations to find groups which minimized differences within and maximized differences without, and when such conduct may have started. It turns out that we identified exactly the set of managers which were suspected of colluding and the year when the practice started. If the same analysis were conducted today, a variety of new and interesting labels might be put on it. This only goes to show that the trend of empirical economics has long been moving towards the principles and methods of "machine learning."

Recall that as a numerical procedure, the techniques of linear regression familiar to the economist are also the techniques of best linear prediction familiar to the data scientist. As a practical matter, both will do exactly the same thing with the data to form their estimates. When an econometrician is conducting a regression analysis, the central question she faces is, what explanatory variables X to include? That is because the central assumption which transforms the best linear predictor of Y from X into a *regression* of Y on X, complete with all the "hypothesis testing" and "statistical significance" arguments, is that all the other factors which influence Y not included in X are uncorrelated with X. In other words, the econometrician has to be comfortable that the "stuff left out" of the analysis is not correlated with the "stuff put in." We use economic theory as our guide to what to include in the regression.

But the variable selection problem is no less important to the data scientist concerned with predicting Y from X. To invoke another tired cliché, "garbage in, garbage out." An economist may be well positioned to identify which variables to include in X to get a useful prediction. Arguably the greater risk is that of over-fitting, that the computer may identify a spurious connection which happens to hold in the training dataset and assume that it will always hold. The best discipline against this over-fitting is the same sort of economic theory the econometrician uses.

This will be true even when we move beyond linear predictors into much more sophisticated classes. A special expertise is quite possibly needed to implement the method, or evaluate which methods are more appropriate for the task at hand, an expertise not always, or even often, found in an economist. But, at the same time, the expertise and judgment about what variables to include, and what form they should take, and what relationships to impose is likely found in an economist more so than a computer scientist.

V. COULD WE LEARN TO TRUST THE MACHINE?

A different question is whether market participants, agencies, and regulatory authorities would ever be able to trust the machine's classification, no matter what diagnostic evidence of its strength could be provided. That is because at its root, from the economist's perspective, an empirical approach to cartel detection is not only a predictive or classification problem: there is usually a testing component, on top of other qualitative considerations. By that we mean that it is more natural, more comfortable and arguably more appropriate to formulate the problem as a hypothesis to be *tested*: how likely is it that the observed data were generated from a collusive rather than a competitive dynamic?

Put this way, this sounds more like an econometric problem than a machine learning problem. What we will need at the end of the day is the ability to parse whether an observed change in price similarities among firms, to take one example, is statistically significant. We do not necessarily need any ability to predict a *future* change in price dispersion. While a machine learned classification algorithm could be enormously helpful in identifying which cases to research further, that further research would still be required, and that research is arguably better suited for an economist.

VI. CONCLUSION

The shoe example above highlights that predicting and understanding are not necessarily the same thing. We can sometimes predict one thing from another without understanding the nature of their relationship. We can also, under the right circumstances, develop an understanding of how one thing relates to another without being in a position to form a useful prediction.

Empirical research into the existence or effectiveness of cartels is best seen as a mixture of both problems. On the one hand, it would be extremely useful to have flexible, powerful, relatively unstructured algorithms which make useful predictions or classifications of whether certain data patterns are anomalous and worthy of further investigation. We fully expect to see an increasing use of machine learning techniques as data on prices, costs, quantities, and fundamental demand and supply drivers become ever more available. The expertise and understanding of economists will be critical to avoid problems of over-fitting and reacting to spurious results.

The second problem will also always remain: the need to formulate the right statistically testable hypothesis. We will want to be able to make statements about the *likelihood* of one hypothesis versus an alternate. Here again the expertise of the economist will prove helpful, though again it may be augmented with the special skills of computational experts.

Empirical analysis has long been moving away from structured assumptions such as assuming things have one probability distribution or another and towards much more data-intensive methods. Kernel regression and isotonic regression are just a couple of examples of completely non-parametric methods which have long been in the econometrics literature. Their popularity has risen as the necessary data requirements are more easily satisfied.

In the end, analysis is best thought of as a continuum. With more data, there is less need to impose structure on the problem. The rapid developments in the theory and practice of artificial intelligence and machine learning are truly exciting. But in our view, they don't replace the need for economic theory and discipline. Instead, they just further expand the toolkit that an economist can bring to bear.