How can machine learning and artificial intelligence be used in development interventions and impact evaluations?

SUBMITTED BY DAVID MCKENZIE (/TEAM/DAVID-MCKENZIE) (HTTPS://TWITTER.COM/INTENT/FOLLOW?SCREEN_NAME=DMCKENZIE001) ON MON, 03/05/2018

2 COMMENTS (HTTP://BLOGS.WORLDBANK.ORG/IMPACTEVALUATIONS/HOW-CAN-MACHINE-LEARNING-AND-ARTIFICIAL-INTELLIGENCE-BE-USED-DEVELOPMENT-INTERVENTIONS-AND-IMPACT#COMMENTS)

Last Thursday I attended a conference on AI and Development organized by CEGA, DIME, and the World Bank's Big Data groups (website (https://www.measuredev.org/), where they will also add video). This followed a World Bank policy research talk last week by Olivier Dupriez on "Machine Learning and the Future of Poverty Prediction" (video (http://bit.ly/2GTsbdO), slides (http://bit.ly/2oyjgHQ)). These events highlighted a lot of fast-emerging work, which I thought, given this blog's focus, I would try to summarize through the lens of thinking about how it might help us in designing development interventions and impact evaluations.

A typical impact evaluation works with a sample *S* to give them a treatment *Treat*, and is interested in estimating something like:
Y(i,t) = b(i,t)*Treat(i,t) +D'X(i,t) for units *i* in the sample *S*
We can think of machine learning and artificial intelligence as possibly affecting every term in this expression:

Measuring outcomes (Y)
One of the biggest use cases currently seems to be in *getting basic measurements* in countries where there are lots of gaps in the basic statistics. Joshua Blumenstock referred to this as "band-aid" statistics. A lot of this work is using either satellite data or cellphone record data to try to predict poverty at a granular level for entire countries or continents (e.g. Oliver's talk (http://bit.ly/2oyjgHQ), Josh's work in Rwanda (http://www.jblumenstock.com/files/papers/jblumenstock_2015_science.pdf), Marshall Burke's work in Africa (https://web.stanford.edu/~mburke/papers/JeanBurkeEtAl2016.pdf)). Other such outcomes being predicted from satellite data include agricultural yields (Marshall Burke's work (https://web.stanford.edu/~mburke/papers/BurkeLobellPNAS2017.pdf)), urbanization (e.g. Ran Goldblatt's work (http://www.mdpi.com/2072-4292/8/8/634)), conflict-affected infrastructure (e.g. Jonathan Hersh's work (http://jonathan-hersh.com/research/)).

At the moment such data seems useful for descriptive work, but it is unclear whether accuracy is enough to **measure changes** well over time – so if you are trying to evaluate the impact of regional or macro policies, there may not be enough signal to be able to detect the impact of interventions, especially over short time horizons. But satellite data are now getting much more accurate, with daily data at relatively high resolution. Christian Clough gave an example of work they are doing in Dar es Salaam, where their challenge has been to detect new buildings going up and changes in building heights to measure where urban growth is taking place. This level of detail could be useful for measuring impacts of transport infrastructure interventions for example.

A second measurement use comes at a more micro level, *enabling measurement of outcomes we might otherwise struggle to measure*. One example comes from work by Ramya Parthasarathy (http://documents.worldbank.org/curated/en/582551498568606865/Deliberative-inequality-a-text-as-data-study-of-Tamil-Nadus-village-assemblies). They use textual analysis of transcripts of India's village assemblies to identify what topics are discussed, and how the flow of conversation varies with gender and status of the speaker. The vast volume of data would make this very hard to do systematically using traditional measurement methods. They can use this to find, for example, that female citizens are less likely to speak, less likely to drive the topic of conversation, and get fewer responses from state officials – but that when the village has been randomly chosen to have a female president, women citizens are more likely to receive a response than with a male president.

A third use case for measurement comes in helping us *decide which outcomes to collect at high frequency.* If we want to do quick surveys that can help us track outcomes over high frequency, machine learning can be used to help determine which subset of variables to collect (e.g. Olivier's talk, Erwin Knippenberg (https://www.erwinknippenberg.com/research/)'s talk).

Targeting the Treatment (selecting S)
A second big use being proposed is to use machine learning to help better target interventions. This can include both *when to intervene* as well as *where/for whom*. Poverty mapping is one obvious example. Other examples given include using remote sensing to detect where deforestation might be starting to take place, to quickly intervene; using machine learning on VAT tax data in India to better target firms for audits (Aprajit Mahajan); predicting travel demand patterns after hurricanes (Scott Farley (https://blog.mapbox.com/tracing-hurricane-marias-wake-7217a7d08380)) or during big events such as the Olympics (Yanyan Xu (http://www.mit.edu/~yanyanxu/doc/Interface_xu_2017.pdf)) to help figure out where transport interventions are needed; predicting where food insecurity will occur to help target aid interventions (Erwin Knippenberg (https://www.erwinknippenberg.com/research/), Daniela Moody (http://www.harrisgeospatial.com/Support/MaintenanceDetail/TabId/3428/ArtMID/13350/ArticleID/23401/Satellite-Imagery-Analysis-for-Automated-Global-Food-Security-Forecasting-%7C-Dr-Daniela-Moody.aspx)); using mobile call records to identify a pool of small businesses that credit can be extended to (Sean Higgins (https://www.seankhiggins.com/research)); and figuring out where there are lots of girls out of school in order for an NGO to figure out which region of India to next expand its program to (Ben Brockman (http://idinsight.org/spotlight-ben-brockman/)).

Almost all of these are in the proof of concept stage right now, showing that such methods *could, in principle*, be used for targeting interventions, but few of them are actually being used by governments to currently target programs. One point that came up in the discussion was that, because some of these methods are quite opaque, policymakers may need a lot of convincing to use them, and may be afraid of the media finding cases where machines have targeted quite wrong. We might therefore expect use to take off first among the private sector. Indeed, Dan Björkegren (http://dan.bjorkegren.com/) noted that one place where it had taken off was in the use of mobile money loans in Kenya, where mobile phone usage data can be used to predict debt repayment, and over 11 million borrowers have now received loans.

AI and ML as part of the treatment (Treat)

There currently seem to be fewer cases where artificial intelligence and machine learning are being used for the interventions themselves, but the promise lies in using them for individualized and dynamic treatments. Jake Kendall outlined a vision for this, noting that his organization have been giving small grants to develop artificial intelligence chatbots that act as digital guides and advocates to help the poor navigate through bureaucracies. Examples included chatbots that could provide immigration help in the Dominican Republic, and help navigate people in the Philippines through a social welfare program. Another example comes from agriculture, where Ofir Reich (https://soundcloud.com/80000-hours/18-ofir-reich-data-science) explained how they were trying to provide customized agricultural advice to farmers through mobile phones, with rapid testing and feedback being used to provide actionable customized information that farmers could use.

Machine learning to measure treatment heterogeneity (b(i,t))

Susan Athey (https://people.stanford.edu/athey/research#econometric) gave an excellent keynote talk that rapidly overviewed how machine learning can be used in economics, and her AEA lectures have more. She noted two different approaches in using machine learning to identify heterogeneity in treatment effects. The first builds on the way we typically do heterogeneity analysis, where we examine heterogeneity by some X variable. The idea here is to use machine learning to figure out what the right groups are for doing so - using causal trees, targeted machine learning, X-learners, or other methods – and then once people are assigned to groups, you can get standard errors on that heterogeneity and it is similar to our standard case. One caveat she noted is in interpreting the groups – e.g. just because the causal tree splits on education and not gender, it does not mean that gender is not important for heterogeneity (the two could be correlated for a start). A second approach is to take a non-parametric approach, and try to get an expected treatment effect for each individual unit. This is what causal forests do. This is a rapidly advancing area, with relatively few practical applications to point to so far. Robert On gave one example – they worked with the One Acre Fund in Rwanda to digitally market lime fertilizer to a massive sample of farmers, and then use this large sample to employ both causal tree and causal forest approaches to examine heterogeneity in treatment impacts.

Taking care of the Confounders (D'X)

In her talk, Susan noted that while machine learning won't solve your identification problem, it can at least help you become more systematic about model selection for the predictive part of your model. This is particularly important in non-experimental applications, and she gave references to machine learning tools for work with matching, instrumental variables, and RDD. Cyrus Samii (https://arxiv.org/abs/1607.03026) provided one example, for work in Colombia where they wanted to examine different policies the government could use to reduce criminality among ex-combatants. Intuitively, selection on observables seems more plausible when you have lots of observables – but with 114 observables, standard OLS or propensity score matching approaches may not work well. His work used regularized propensity score methods and compared them to these other approaches – yielding estimates of the impacts of employment and socio-emotional support programs.

Challenges and Reflections

A few final notes of some of the key challenges and areas for future work:

1. **What is the gold standard?** Supervised machine learning requires a labeled training data set and a metric for evaluating performance. This raises several challenges. The first is that the very lack of data that these approaches are trying to solve also makes it hard to train the data in the first place. As a result, researchers have often had to collect a lot of survey data or get people to hand-label images in order to have something to train against. A second challenge is that survey data is not error-free – so if you predict someone to be poor, but the survey says they aren't, it isn't clear which is the error. Sol Hsiang (https://gspp.berkeley.edu/directories/faculty/solomon-hsiang#research) discussed one potential approach to this problem – develop the models in an environment where you have really great data (e.g. the U.S.), and then start degrading the data to see how the model would perform under developing country data conditions – an approach that still needs validating with actual developing country applications.
2. **Beware of the hype/are we learning about enough failures?** I discussed my work on trying to predict successful entrepreneurs (http://blogs.worldbank.org/impactevaluations/can-predicting-successful-entrepreneurship-go-beyond-choose-smart-guys-their-30s-comparing-machine), for which machine learning did not do very well. But this was the only case of failure I saw out of 25+ presentations – surely the failure rate is much higher than 4%! While many presenters were appropriately cautious, there was also a high ratio of pretty pictures to demonstrated impact. We need to be better about also making clear when these methods do not offer improvements (or when they do worse) than current methods.
3. **Dealing with dynamics:** I) a first concern is how stable many of the predicted relationships are. That is, if conduct an expensive training set survey to help me predict the relationship between satellite images and crop yields today, will this same relationship still hold in a year's time, or 5 year's time? II) a second concern is that of behavioral responses - e.g. if people learn their phone calling behavior is being used to determine eligibility for interventions, they may change their behavior. Apparently there is something called "adversarial machine learning" that is a frontier topic to think about designing methods more robust to this.
4. **Ethics/Privacy/Fairness** – lots of issues here – is it fair to be denied a program because the people you talk to on your cellphone have really variable calling patterns? What rights do people have to privacy in an environment where satellites are photographing their house every day, phones are tracking their every move and communication, their moods are being analyzed on social media, etc.? And given all these concerns, how much will access to this type of data become the preserve of a very limited subset of researchers?

Apologies to anyone whose work I misrepresented or that didn't fit within the lens I chose for summarizing. Feel free to add better links to your work too below. I welcome any comments, especially from those who want to share lessons from failures…

Comments

Hi David (/impactevaluations/comment/4766#comment-4766)

SUBMITTED BY RICK DAVIES ON MON, 03/05/2018 - 08:16

Hi David

A fourth challenge: how to make machine learning more user-friendly, and thus used more widely.

One option is to use EvalC3 - a free Excel app (https://evalc3.net/) that can be used for prediction modeling with small data sets, which in my experience of development aid projects are far more common than large data sets. It includes both manual model design and model design using algorithms (decision trees, exhaustive search and genetic algorithm), with mutiple measures of model performance. The workflow also includes case selection tools for within-case inquiry into any causal mechanisms at work. PS: Decision Tree models by definition allow for heterogeneity aka equifinality.

More sophisticated options include BigML and Rapid Miner Studio, both with free usage options.

And have you read "Bit by Bit: Social Research in the Digital Age" by Mathew Salganik, published this year? Very readable and inspiring

regards, rick davies
www.mande.co.uk

reply (/impactevaluations/comment/reply/1648/4766)

For a simple example of (/impactevaluations/comment/4769#comment-4769)

SUBMITTED BY RICK DAVIES ON MON, 03/05/2018 - 18:00

For a simple example of predictive modeling of poverty status, using Rapid Miner Studio and an existing data set from 2006 rural Vietnam see http://mande.co.uk/special-issues/the-basic-necessities-survey/#simple

regards, rick davies

reply (/impactevaluations/comment/reply/1648/4769)

Add new comment

Your name

E-mail

The content of this field is kept private and will not be shown publicly.

Comment *

More information about text formats (/impactevaluations/filter/tips)

Allowed HTML tags:
 <p>
Lines and paragraphs break automatically.

By submitting this form, you accept the Mollom privacy policy (https://www.mollom.com/web-service-privacy-policy).

Save Preview