Pre-read: ads and recommender systems

COS 597E/SOC 555, Princeton University, Fall 2020

There is a popular narrative that machine learning excels at predicting which ads people will click on, and that tech companies have made trillions of dollars by doing this well. You may have seen a quote by data scientist Jeff Hammerbacher: “The best minds of my generation are thinking about how to make people click ads. That sucks.”

This week we’ll learn just how oversimplified this narrative is, and what that tells us about the limits to prediction and the limits of prediction. Specifically:

- Problem definition: click prediction is only one small part of the problem of maximizing the effectiveness of advertising. It is unclear what role (if any) machine learning has in the other aspects of the problem.
- Stakeholder perspectives: the problem formulation from the ad platform’s perspective is different from the advertiser’s perspective, which is in turn different from the publisher’s perspective and the user’s perspective.
- Different types of limits: as usual, there are scientific limits to predictive accuracy and challenges in measuring predictive accuracy. But there are also computational limits, engineering limits, commercial limits, and normative limits (e.g. privacy).

Here’s a simplified formulation of the click prediction problem: given the content of the ad and the context in which the ad appears, predict the probability that the ad will be clicked on (based on past click data). In search advertising, the context is the query that the user typed. In display advertising, the context may be the web page on which the ad appears (among other things). The above formulation corresponds to non-personalized advertising. In personalized advertising, the click probability depends not just on the ad content and the context but also on knowledge of the user and her past activities. If you want a primer on online advertising, see this article; specifically, the section “Review of ads: display, branded, targeted, and programmatic”.

Our first reading is a 2013 paper by Google engineers/researchers. There are two notable things for our purposes: first, even with Google’s computational resources, the scale of the data is such that they make a number of tradeoffs which may potentially impact accuracy (Sections 2--4).[1] Second, measuring accuracy is a tricky problem as usual (Sections 5--7). There are some familiar issues such as absolute vs relative predictability, but also some new ones such as feedback loops, which are pervasive with Internet algorithms. That is, the predictions generated by the model during one time period become part of the input during the next time period.[2]

The Google paper casually drops the concept of Bayes error (which is becoming a common pattern) but doesn’t offer any concrete ideas as to what the limits to prediction are. In fact, it doesn’t even report the Click Through Rate (CTR) that the system achieved! This is because CTR values tend to be closely guarded industry secrets. The CTR for display advertising generally tends to be around 0.1% --- that is, a thousandth of ads seen are clicked on. At first sight that seems low, but how can we contextualize this? What would be the CTR of a hypothetical perfect ad?

[1] Although this paper is 7 years old, it is unlikely that things are qualitatively different today considering that the scale of the data has grown, and not just hardware capabilities.

[2] We saw an example of a feedback loop in the computer vision topic: Recht et al. couldn’t use Flickr’s ranking features to retrieve images because those features may rely on labels generated by training on ImageNet.

We can also look at the CTR of personalized advertising relative to non-personalized advertising. Note that the Google paper is about non-personalized advertising. There are different ways to measure the "lift" in CTR that personalization achieves, but one recent industry estimate (also by Google) puts it at around 50%. Again, is this "high" or "low"? Here's one way to think about it. Ad personalization can be seen as the problem of selecting the users that are the best targets for a given ad (rather than the best ads for a given user). Consider an ad for a software engineering job. Even if the position is fully remote, the ad is relevant to maybe 1% of users who are interested in software engineering jobs. If the personalization algorithm manages to find those users, it can potentially have a 100x higher (rather than a 1.5x higher) CTR than a non-personalized ad. This thought experiment omits many nuances, but it is a useful starting point for discussion.

To summarize, it is unclear if there is a fundamental reason why the CTR of online ads, especially personalized ads, couldn't be 10x higher or even 100x higher than it is today. What would be the commercial and societal implications of a world in which ads were a hundred times more effective? Do you think ads will become much more effective if data collection becomes even more pervasive and computational limits go away? Or are there limits we haven't discovered or articulated?

Last week we discussed the difference between prediction and action. That distinction is crucial here as well. So far we've only talked about prediction. Given a system for CTR prediction, how does the ad platform use it to figure out which ads to show? We've loosely assumed that the goal is maximizing the CTR, but that's not really true. The expected revenue is the product of click probability and cost per click.[3] This is what the company is trying to maximize --- in aggregate, rather than for a single ad slot. This introduces many complications, chiefly the need to figure out the cost per click. Recall that these costs have to be determined in an automated way!

This is the problem of designing ad auctions, studied in the field of mechanism design. If you want to learn more, here is a lecture/chapter that introduces auction design, explains some of the core ideas of the field, and applies them to ad auctions (it is not one of our assigned readings). You may notice that the whole thing seems orthogonal to prediction and machine learning. Indeed, there is an entire field of people working on the problem of serving ads who don't think about predicting clicks.

So far, while we've broadened the problem beyond click prediction, we're still within the ad platform's perspective. Two of the other players in this business are the advertiser and the publisher, and there is of course the user.[4] Let's consider the advertiser, and defer a discussion of the publisher's and user's perspectives to the class meeting.

The advertiser's problem looks nothing like the prediction/optimization/matching problem that the ad platform tries to solve. The advertiser ultimately cares about getting people to buy its product. That means it would be useless to advertise to people who were going to buy the product anyway. What the advertiser needs to do is maximize the *additional* purchases *because* of the ad. That's a causal inference problem. This is another theme in this course: one reason why even accurate prediction may not be useful is because what we actually need is causal inference. Doing causal inference right is extremely tricky, and vanilla machine learning (without a domain-specific theoretical framework) is not a good tool for it.

[3] Our discussion assumes that the advertiser is charged per click, but they could also be charged per impression, i.e. based on the number of times the ad is viewed.

[4] There are many other players in the ad business that we won't consider. There's a famous chart that shows how complex it is.

The Lewis, Rao, and Reiley paper explains how hard this problem is: (1) since such a low fraction of people exposed to an ad end up buying the product, even with Randomized Controlled Trials --- the gold standard of causal inference --- enormous sample sizes are needed to measure the effect of an ad (2) if you're trying to measure the effect of advertising from purely observational data --- the far more common approach in the industry --- even the minutest selection bias poses a problem (3) ad personalization is precisely a vehicle for selection bias, as it targets ads to people who are already more likely to buy the product; in typical settings, its effect will be far larger than the impact of seeing the ad and (4) CTR is a poor proxy for the things that matter to the advertiser.

This is an economics paper. Even though its main argument is a simple one, it uses terms that may be unfamiliar to many of you. If you find this paper difficult, feel free to skim it and read the thoroughly entertaining general-audience article that is based on this and other papers. It makes the case that even though trillions are spent on advertising every year, there is currently no effective way to know if any of it is working.

For Thursday, we will turn from ads to recommender systems. These algorithms suggest news articles, books, movies, videos, music, and products online, and are considered another major commercial success story of machine learning. There are many similarities and differences between ads and recommendations.[5] The high-level observations we made about ads are true of recommendations as well: click prediction is just a small part of the overall problem; the incentives of users, creators, and the platform are not fully aligned; and there are many types of limits to prediction: scientific, computational, engineering, etc. Our readings will reinforce these points and also contain a case study of how things can go wrong if we ignore them.

We'll start with a 2004 paper that explains in detail why recommendation is not just a matter of maximizing predictive accuracy and, even to the extent that it is, there isn't one single measure that's always appropriate. It presents dozens of goals and metrics (and in retrospect there are even more). It also presents tentative evidence that the metrics for predictive accuracy that they considered cluster into three broad groups.

Next we turn to the Netflix Prize, a million dollar machine learning contest to predict users' ratings on movies (evaluated on a held-out set). It was announced in 2006 to great fanfare and attracted tens of thousands of contestants (see here for a bit of background and one contestant's perspective if you're interested in learning more). But had a crude prediction problem formulation and a single evaluation metric. Pretty soon, contestants observed many things about the data that suggested strong limits to predictability. For example, multiple members of a household with different tastes may share an account. If we don't know (and can't figure out) which member rated a movie, predictive accuracy will be lower. Netflix almost surely knew of these limits: after all, the target that Netflix set for winning the $1 million prize corresponded to an improvement in predictive accuracy equivalent to a mere 0.2 stars out of 5 compared to simply predicting the mean — for example, predicting a 3.2 instead of a 3.0 when the customer in fact rated the movie a 4.0. However, the hope seems to have been that the scientific insights obtained from the problem setting would generalize to the production setting.

What in fact happened is that most of the seeming scientific progress from the contest was a case of "overfitting to the problem formulation". This is another recurring theme but one that gets relatively little attention. Specifically, the two main insights behind the winning solutions were matrix factorization and

[5] One industry maxim is that a recommendation is just a really effective ad. Can you think of formats that attempt to blur the line?

ensembles (we'll discuss these in class). They are both excellent ideas, but when we take into account how Netflix's actual goals and constraints differ from the competition setting, neither proved useful.

Indeed, Netflix did not adopt the solutions that resulted from the competition, as explained in the next reading. There are even more practical considerations not mentioned in this blog post which we will discuss in class. Also worth noting: the reading takes the simplistic view that the company is simply trying to do what's good for the user, but in fact there is an important misalignment of incentives between users and platforms, which has become a flashpoint in criticisms of Facebook's and YouTube's recommendation systems.

The Netflix Prize was as significant for recommender systems as ImageNet was for computer vision in terms of the attention paid to them in the respective communities. But the impacts were different. For example, we discussed how the use of a single evaluation metric for each task was not a limitation in the case of ImageNet, but it seems to have limited progress in the case of the Netflix Prize. What do you think explains the difference?

Our final reading is a paper with a cute title that is all about practical considerations. Specifically, it explains how machine learning clashes with the principles of sound software engineering. While our main interest in this course is the scientific limits to prediction, these practical aspects are good to keep in mind. Note that this was written by some of the same Google authors as the "View from the trenches" paper.