How Cambridge Analytica's Facebook targeting model really worked – according to the person who built it

March 30 2018, by Matthew Hindman

Credit: AI-generated image (disclaimer)

The researcher whose work is at the center of the Facebook-Cambridge Analytica data analysis and political advertising uproar has revealed that his method worked much like the one Netflix uses to recommend movies.

In an email to me, Cambridge University scholar Aleksandr Kogan explained how his statistical model processed Facebook data for Cambridge Analytica. The accuracy he claims suggests it works about as well as established voter-targeting methods based on demographics like race, age and gender.

If confirmed, Kogan's account would mean the digital modeling Cambridge Analytica used was hardly the virtual crystal ball a few have claimed. Yet the numbers Kogan provides also show what is – and isn't – actually possible by combining personal data with machine learning for political ends.

Regarding one key public concern, though, Kogan's numbers suggest that information on users' personalities or "psychographics" was just a modest part of how the model targeted citizens. It was not a personality model strictly speaking, but rather one that boiled down demographics, social influences, personality and everything else into a big correlated lump. This soak-up-all-the-correlation-and-call-it-personality approach seems to have created a valuable campaign tool, even if the product being sold wasn't quite as it was billed.

The promise of personality targeting

In the wake of the revelations that Trump campaign consultants Cambridge Analytica used data from 50 million Facebook users to target digital political advertising during the 2016 U.S. presidential election, Facebook has lost billions in stock market value, governments on both sides of the Atlantic have opened investigations, and a nascent social movement is calling on users to #DeleteFacebook.

But a key question has remained unanswered: Was Cambridge Analytica really able to effectively target campaign messages to citizens based on their personality characteristics – or even their "inner demons," as a

company whistleblower alleged?

If anyone would know what Cambridge Analytica did with its massive trove of Facebook data, it would be Aleksandr Kogan and Joseph Chancellor. It was their startup Global Science Research that collected profile information from 270,000 Facebook users and tens of millions of their friends using a personality test app called "thisisyourdigitallife."

Part of my own research focuses on understanding machine learning methods, and my forthcoming book discusses how digital firms use recommendation models to build audiences. I had a hunch about how Kogan and Chancellor's model worked.

So I emailed Kogan to ask. Kogan is still a researcher at Cambridge University; his collaborator Chancellor now works at Facebook. In a remarkable display of academic courtesy, Kogan answered.

His response requires some unpacking, and some background.

From the Netflix Prize to "psychometrics"

Back in 2006, when it was still a DVD-by-mail company, Netflix offered a reward of $1 million to anyone who developed a better way to make predictions about users' movie rankings than the company already had. A surprise top competitor was an independent software developer using the pseudonym Simon Funk, whose basic approach was ultimately incorporated into all the top teams' entries. Funk adapted a technique called "singular value decomposition," condensing users' ratings of movies into a series of factors or components – essentially a set of inferred categories, ranked by importance. As Funk explained in a blog post, "So, for instance, a category might represent action movies, with movies with a lot of action at the top, and slow movies at the bottom, and correspondingly users who like action movies at the top, and those who

prefer slow movies at the bottom."

Factors are artificial categories, which are not always like the kind of categories humans would come up with. The most important factor in Funk's early Netflix model was defined by users who loved films like "Pearl Harbor" and "The Wedding Planner" while also hating movies like "Lost in Translation" or "Eternal Sunshine of the Spotless Mind." His model showed how machine learning can find correlations among groups of people, and groups of movies, that humans themselves would never spot.

Funk's general approach used the 50 or 100 most important factors for both users and movies to make a decent guess at how every user would rate every movie. This method, often called dimensionality reduction or matrix factorization, was not new. Political science researchers had shown that similar techniques using roll-call vote data could predict the votes of members of Congress with 90 percent accuracy. In psychology the "Big Five" model had also been used to predict behavior by clustering together personality questions that tended to be answered similarly.

Still, Funk's model was a big advance: It allowed the technique to work well with huge data sets, even those with lots of missing data – like the Netflix dataset, where a typical user rated only few dozen films out of the thousands in the company's library. More than a decade after the Netflix Prize contest ended, SVD-based methods, or related models for implicit data, are still the tool of choice for many websites to predict what users will read, watch, or buy.

These models can predict other things, too.

Facebook knows if you are a Republican

In 2013, Cambridge University researchers Michal Kosinski, David Stillwell and Thore Graepel published an article on the predictive power of Facebook data, using information gathered through an online personality test. Their initial analysis was nearly identical to that used on the Netflix Prize, using SVD to categorize both users and things they "liked" into the top 100 factors.

The paper showed that a factor model made with users' Facebook "likes" alone was 95 percent accurate at distinguishing between black and white respondents, 93 percent accurate at distinguishing men from women, and 88 percent accurate at distinguishing people who identified as gay men from men who identified as straight. It could even correctly distinguish Republicans from Democrats 85 percent of the time. It was also useful, though not as accurate, for predicting users' scores on the "Big Five" personality test.

There was public outcry in response; within weeks Facebook had made users' likes private by default.

Kogan and Chancellor, also Cambridge University researchers at the time, were starting to use Facebook data for election targeting as part of a collaboration with Cambridge Analytica's parent firm SCL. Kogan invited Kosinski and Stillwell to join his project, but it didn't work out. Kosinski reportedly suspected Kogan and Chancellor might have reverse-engineered the Facebook "likes" model for Cambridge Analytica. Kogan denied this, saying his project "built all our models using our own data, collected using our own software."

What did Kogan and Chancellor actually do?

As I followed the developments in the story, it became clear Kogan and Chancellor had indeed collected plenty of their own data through the thisisyourdigitallife app. They certainly could have built a predictive

SVD model like that featured in Kosinski and Stillwell's published research.

So I emailed Kogan to ask if that was what he had done. Somewhat to my surprise, he wrote back.

"We didn't exactly use SVD," he wrote, noting that SVD can struggle when some users have many more "likes" than others. Instead, Kogan explained, "The technique was something we actually developed ourselves … It's not something that is in the public domain." Without going into details, Kogan described their method as "a multi-step co-occurrence approach."

However, his message went on to confirm that his approach was indeed similar to SVD or other matrix factorization methods, like in the Netflix Prize competition, and the Kosinki-Stillwell-Graepel Facebook model. Dimensionality reduction of Facebook data was the core of his model.

How accurate was it?

Kogan suggested the exact model used doesn't matter much, though – what matters is the accuracy of its predictions. According to Kogan, the "correlation between predicted and actual scores … was around [30 percent] for all the personality dimensions." By comparison, a person's previous Big Five scores are about 70 to 80 percent accurate in predicting their scores when they retake the test.

Kogan's accuracy claims cannot be independently verified, of course. And anyone in the midst of such a high-profile scandal might have incentive to understate his or her contribution. In his appearance on CNN, Kogan explained to a increasingly incredulous Anderson Cooper that, in fact, the models had actually not worked very well.

In fact, the accuracy Kogan claims seems a bit low, but plausible. Kosinski, Stillwell and Graepel reported comparable or slightly better results, as have several other academic studies using digital footprints to predict personality (though some of those studies had more data than just Facebook "likes"). It is surprising that Kogan and Chancellor would go to the trouble of designing their own proprietary model if off-the-shelf solutions would seem to be just as accurate.

Importantly, though, the model's accuracy on personality scores allows comparisons of Kogan's results with other research. Published models with equivalent accuracy in predicting personality are all much more accurate at guessing demographics and political variables.

For instance, the similar Kosinski-Stillwell-Graepel SVD model was 85 percent accurate in guessing party affiliation, even without using any profile information other than likes. Kogan's model had similar or better accuracy. Adding even a small amount of information about friends or users' demographics would likely boost this accuracy above 90 percent. Guesses about gender, race, sexual orientation and other characteristics would probably be more than 90 percent accurate too.

Critically, these guesses would be especially good for the most active Facebook users – the people the model was primarily used to target. Users with less activity to analyze are likely not on Facebook much anyway.

When psychographics is mostly demographics

Knowing how the model is built helps explain Cambridge Analytica's apparently contradictory statements about the role – or lack thereof – that personality profiling and psychographics played in its modeling. They're all technically consistent with what Kogan describes.

A model like Kogan's would give estimates for every variable available on any group of users. That means it would automatically estimate the Big Five personality scores for every voter. But these personality scores are the output of the model, not the input. All the model knows is that certain Facebook likes, and certain users, tend to be grouped together.

With this model, Cambridge Analytica could say that it was identifying people with low openness to experience and high neuroticism. But the same model, with the exact same predictions for every user, could just as accurately claim to be identifying less educated older Republican men.

Kogan's information also helps clarify the confusion about whether Cambridge Analytica actually deleted its trove of Facebook data, when models built from the data seem to still be circulating, and even being developed further.

The whole point of a dimension reduction model is to mathematically represent the data in simpler form. It's as if Cambridge Analytica took a very high-resolution photograph, resized it to be smaller, and then deleted the original. The photo still exists – and as long as Cambridge Analytica's models exist, the data effectively does too.

This article was originally published on The Conversation. Read the original article.

Provided by The Conversation

Citation: How Cambridge Analytica's Facebook targeting model really worked – according to the person who built it (2018, March 30) retrieved 4 May 2024 from https://techxplore.com/news/2018-03-cambridge-analytica-facebook-person-built.html