Criteo Uplift Modelling – Project Report

Adriaan van Heerden | github.com/avanheerden/criteo-uplift-modelling

Background & Dataset

Criteo (criteo.com) is a digital advertising company – the kind that shows you a banner ad for a pair of shoes you looked at once and then follows you around the internet for a week. In 2018 they published a research paper based on a large marketing experiment across 13 million users and made the accompanying dataset publicly available for data scientists to study.

In the experiment, users were randomly split into two groups. The **treatment group** (about 85% of users) were shown Criteo's advertisements. The **control group** (the remaining 15%) were not. Criteo then tracked whether each user visited the advertiser's website and whether they made a purchase.

The dataset records twelve pieces of information about each user, though the exact nature of these features has been anonymised. It also records whether the user was in the treatment or control group, whether they actually saw the ad, whether they visited the site, and whether they bought something.

Field	Description
f0 – f11	Twelve anonymised numerical features representing user characteristics (e.g. purchase history, browsing behaviour, demographics)
treatment	Whether the user was assigned to the treatment group (1) or control group (0) – randomly assigned
exposure	Whether the user actually saw the advertisement (1) or not (0) – being assigned to treatment doesn't guarantee exposure
visit	Whether the user subsequently visited the advertiser's website – **primary outcome**
conversion	Whether the user made a purchase – a secondary outcome, rarer and harder to model reliably

In this dataset, a “treated user” is someone who was **selected to receive the campaign** – they were entered into the advertising system as a target. Whether the ad was actually delivered to them is a separate question, captured by the exposure variable. So the terminology maps as follows:

- **Treated user** (treatment = 1) – selected as a campaign target; 11.9 million users
- **Exposed user** (exposure = 1) – actually saw the advertisement; 428,000 users (~3.6% of treated users)
- **Control user** (treatment = 0) – deliberately withheld from the campaign; 2.1 million users

The problem

When a company spends money on advertising, the obvious question is: did it work? Did people buy things because of the ad, or would they have bought them anyway?

This turns out to be a surprisingly tricky question. Imagine you run an email campaign and 10% of recipients make a purchase. That sounds good, but if 9% of people who *didn't* receive the email also made a purchase, the campaign only drove 1 percentage point of additional sales. The other 9% were going to buy regardless.

Now imagine the same campaign, but this time you can identify in advance which users fall into which category:

- **Persuadables** – people who will buy *because* of the campaign, and wouldn't have otherwise
- **Sure Things** – people who will buy regardless (contacting them wastes money)
- **Lost Causes** – people who won't buy regardless (contacting them also wastes money)
- **Sleeping Dogs** – people who would have bought, but are actually put off by being contacted (the most dangerous group)

A standard approach to marketing analytics tries to predict *who is most likely to buy*. But this mostly finds Sure Things: people who were going to buy anyway. What you actually want to find is the Persuadables. That is what **uplift modelling** is designed to do: predict not who will buy, but who will buy *because of* the campaign.

We can estimate the approximate size of each segment from the dataset, though with an important caveat. Because we can never observe what would have happened to a user under the alternative condition – a treated user (someone who was selected to receive the campaign) who visited might or might not have visited without the ad – the segment sizes cannot be measured directly. They can only be estimated by comparing response rates between the treatment and control groups, and only under the assumption that there are no Sleeping Dogs (that the campaign can help or have no effect, but cannot actively harm). This is a simplifying assumption that may not hold perfectly in practice. With that caveat noted, the data suggest the following approximate breakdown across the full 13 million users in the dataset:

Segment	Estimated Proportion	Estimated Users
Persuadables	1.0%	~144,000
Sure Things	3.8%	~534,000
Lost Causes	95.2%	~13,301,000
Sleeping Dogs	0% (assumed)	0

The most striking feature of this breakdown is how small the Persuadable population is: roughly 1 in 100 users. The vast majority of users fall into the Lost Causes category: they will not visit or purchase regardless of whether they see the ad. This is typical for digital advertising at scale, but it underlines why efficient targeting matters so much: broadcasting a campaign to the full population means spending around 95% of the budget on users it cannot influence.

The approach

The analysis was structured as three separate stages.

Stage 1: Is the experiment valid?

Before drawing any conclusions from the data, it is essential to verify that the experiment was run correctly. The most important check is whether the two groups – treatment and control – were genuinely comparable before the campaign started. If the treatment group happened to contain more frequent shoppers than the control group, any difference in outcomes could reflect that pre-existing difference rather than the effect of the campaign.

This was tested using a statistical technique called Standardised Mean Difference, which measures how similar the two groups are across all twelve user features. If the experiment was properly randomised, the groups should look essentially identical before the campaign begins.

Stage 2: Did the campaign work on average?

The second stage measured the overall effect of the campaign – not for individual users, but across the population as a whole. This is a standard A/B test: compare the visit and purchase rates between the **treatment group** (users selected to receive the campaign) and the **control group** (users deliberately withheld from it) and determine whether any difference is large enough to be confident it wasn't just chance.

This stage also examined the practical size of the effect, not just whether it was statistically detectable. With 13 million users, even a trivially small difference will appear statistically significant – so it is important to ask whether the effect is large enough to matter commercially, not just whether it exists at all.

Stage 3: Who does the campaign work for?

The third stage built machine learning models to predict the individual-level effect of the campaign: which specific users were most likely to respond *because* of the ad. Three different model types were trained and compared.

Each model was evaluated using a **Qini curve**: a chart that shows, if you target the users the model ranks most highly, how much of the total incremental benefit you capture. A model that perfectly identifies Persuadables would capture all the benefit by targeting a small fraction of users. A random selection of users would capture benefit in proportion to how many you contact.

The findings

The experiment was valid

The randomisation worked correctly: the treatment and control groups were statistically indistinguishable across all twelve user features before the campaign ran. This means the results can be trusted: any differences in outcomes between the groups were caused by the campaign, not by pre-existing differences between the users.

The campaign worked – but less than it appears

Overall, treated users visited the site at a rate of 4.85%, compared to 3.82% in the control group, i.e. a difference of about 1 percentage point, or a 27% relative improvement. For purchases, the treated group converted at 0.31% versus 0.19% in the control group: a 59% relative improvement.

Both effects were statistically significant, but the effect sizes were small in absolute terms. The campaign moved the needle, but modestly.

The real story is in the exposure data

Here is where it gets interesting. Being assigned to the treatment group did not mean a user actually saw the advertisement: only about 3.6% of users in the treatment group were actually exposed to the ad. The other 96.4% were targeted but never reached.

When the analysis was restricted to users who actually saw the ad, the picture changed dramatically. Among exposed users, the visit rate was 41%, compared to 3.8% in the control group. Among users who were targeted but never saw the ad, the visit rate was actually *lower* than the control group. There are two possible reasons for this:

- **Composition** – The exposed users (those 3.6% who actually saw the ad) have a visit rate of 41%. They are almost certainly self-selected for high commercial intent: people who noticed and engaged with an ad are by definition more actively browsing, more interested in the product category, and more likely to visit regardless. When you remove this high-intent subgroup from the treatment pool, the remaining unexposed users are left with a visit rate slightly *below* the overall average; and since control is essentially the population average, the unexposed treated group ends up just below it. In other words, the unexposed treated group is the treatment group minus its most commercially active members. That subtraction is enough to push the rate marginally below control.

- **Displacement** – Some users who would have visited organically (by typing the URL directly or clicking a search result) may instead have visited by clicking the ad, which counts them as exposed. These organic-intent visits are therefore reclassified into the exposed bucket, slightly deflating the unexposed visit rate.

This means the campaign's average effect – that modest 1 percentage point lift – is almost entirely being dragged down by the 96% of targeted users who were never reached. When the ad actually gets in front of someone, it is highly effective. The problem is that it almost never does.

The dataset does not explain *why* so few targeted users saw the ad, but in digital advertising several factors commonly contribute to this kind of gap between targeting and reach:

- **Ad blockers** – a significant proportion of internet users run software that prevents ads from being served or displayed at all
- **Auction mechanics** – in programmatic advertising, being "targeted" means being entered into a real-time auction for ad inventory; winning that auction depends on bid price, competition from other advertisers, and available inventory. A user can be targeted but never win an auction slot during the campaign window
- **Viewability** – an ad can be technically served but appear below the fold of a page the user never scrolls to; depending on how Criteo defined exposure, some served ads may not have counted as genuinely seen
- **Session timing** – users who were targeted may simply not have been browsing in contexts where the ad could be shown during the campaign period
- **Frequency capping** – campaigns often limit how many times a user can be shown an ad, but the inverse problem (users who are targeted but never reached at all) is common when campaign budgets are insufficient to achieve broad reach across the full targeted audience

It is also worth noting that without a benchmark for comparable campaigns, we cannot say whether 3.6% is unusually low or typical for this type of advertising at this scale. The data tell us that the gap exists and that it matters enormously for outcomes, but it does not allow us to conclude that the campaign was poorly run.

The most commercially valuable finding from the entire analysis is therefore not about whom to target, but about delivery: improving the rate at which the ad actually reaches the people it is aimed at would have far more impact than any refinement to the targeting strategy itself.

It is also worth noting that visits and purchases will not scale proportionally with improved reach. Among exposed users who visited the site, approximately 13% went on to make a purchase, compared to approximately 5% among control visitors. This suggests the ad is reaching users with genuinely stronger purchase intent, not just higher browsing activity. However, the majority of users the ad reaches visit without purchasing, meaning improved delivery should be paired with on-site conversion optimisation to fully realise the commercial value of additional reach.

The uplift models were useful but not dramatically different from each other

Three machine learning models were trained to predict which users would respond most strongly to the campaign: Logistic Regression (the simplest approach), Random Forest, and XGBoost (the most sophisticated).

All three outperformed random targeting, confirming that the user features do contain genuine information about who is more persuadable. Targeting the top 10% of users as ranked by the best model captured 87% of the total incremental benefit from the campaign, compared to the 10% you would expect from random selection.

However, the three models were essentially indistinguishable in their performance. The additional complexity of XGBoost produced no meaningful improvement over simple Logistic Regression. In a real deployment, the simpler model would be the sensible choice: it is faster, easier to maintain, and just as accurate.

The most predictive model was the one focused on exposure

A second set of models was trained asking a different question: not “who should we target?” but “among users who see the ad, who will respond most strongly?” This model – trained on exposed users versus the control group – was 30 times more accurate than the targeting model.

This tells us something important about the data. The twelve user features are good at predicting how someone will respond *once they see the ad*, but they are poor predictors of whether the ad will actually reach them in the first place. Whether an ad is served and viewed depends heavily on platform mechanics – bid prices, inventory, timing – that have nothing to do with the user's profile. The features describe the person; they do not describe whether the advertising platform will successfully show them the ad.

What this means in practice

For a business running a campaign like this one, the analysis points to three actionable conclusions:

1. **Fix the delivery problem first.** With only 3.6% of targeted users seeing the ad, even a modest improvement in reach would generate far more incremental visits and purchases than any improvement in who is targeted.

2. **Use the uplift model to prioritise, not to segment.** The model cannot cleanly separate Persuadables from everyone else, but it can rank users meaningfully. Concentrating budget on the highest-ranked users is genuinely more efficient than targeting randomly. Here is how the ranking works in practice. For each user, the model takes their twelve feature values and predicts the probability of belonging to each of the four target classes: CN (control, no visit), CR (control, visited), TN (treated, no visit), and TR (treated, visited). It then combines these four probabilities into a single uplift score using a formula that, in plain terms, asks: given this user's features, how much more likely are they to visit *because of the treatment* compared to a typical user?

   A user with a high uplift score is one whose feature profile (browsing behaviour, purchase history, and so on) resembles the profile of users who responded positively to the campaign in the training data. A user with a low uplift score resembles users who did not respond, or who visited regardless of treatment. The model does not know for certain which category any individual belongs to – it cannot, because we can never observe both outcomes for the same person – but across a large population, users with higher scores will contain a greater concentration of Persuadables than users with lower scores.

   The practical application is straightforward: sort the entire user population by uplift score from highest to lowest, and target the top n%, where n is determined by the campaign budget. As the targeting efficiency table above shows, targeting the top 10% of users captures 87% of the total incremental benefit, at 10% of the cost of targeting everyone. The model does not need to perfectly identify Persuadables to be useful, it just needs to concentrate them toward the top of the ranking, which it demonstrably does. It is also worth noting that the data contains a weak signal that Sleeping Dogs may exist in this population: users who were targeted but never saw the ad performed slightly below the control group on both visit and purchase rates. The uplift model cannot cleanly identify this segment from the available features, but it is a reason to avoid blanket targeting even when reach improves — some users may respond negatively to contact.

   To illustrate the scale of this efficiency gain, consider a campaign using a standard programmatic display advertising rate of $2 per thousand impressions (a conservative mid-range figure for this type of advertising). The comparison between blanket targeting and uplift-guided targeting looks like this:

Approach	Users targeted	Estimated cost	Incremental visits captured	Cost per incremental visit
Target all treated users (as in dataset)	11,882,655	~$23,765	~122,885 (100%)	~$0.19
Target top 10% by uplift score	1,188,266	~$2,377	~106,664 (87%)	~$0.02

Targeting the top 10% of users costs 90% less and still captures 87% of the total incremental benefit: roughly a 10x improvement in cost efficiency. The cost-per-incremental-visit falls from $0.19 to $0.02.

Two caveats apply. First, the $2 CPM figure is an industry assumption. The dataset contains no actual cost data, so these numbers are illustrative rather than precise. Second, because all users receive positive uplift scores from the model (a limitation discussed in the technical notebooks), the 10% figure reflects the highest-ranked users by *degree* of persuadability rather than a cleanly identified Persuadable segment. The efficiency gain is real, but the top 10% will contain some Sure Things and Lost Causes alongside the Persuadables.

3. **Do not over-engineer the model.** The simplest model performed as well as the most complex one. In data science, a result that holds up under the simplest approach is often a more reliable result than one that requires sophisticated machinery to detect.