Sudharshan
Nov 9, 2020 · 8 min read · Listen

Netflix A/B Testing

(aka confounding the perfect Randomized Controlled Experiment can be a pipe dream)

I was heavily inspired by the Netflix Technology Blog's article on how they go about conducting A/B Tests on movie/series thumbnails — customizing thumbnails according to the user's demographics/preferences to improve the likelihood of them watching a show.

It goes into great depth about the experiment design, but one of the most interesting insights was around Sense 8. Their insights concluded that US audience loved looking at image $\frac{1}{\text{Top Left}}$ — body simple (Top Right image, with one character right in centre, Bottom Left image, no text or title) whereas Germany audience loves looking at abstract images (Top Left image)

This inspired me to go find answers to a question that always puzzled me as an 'International' viewer who primarily watched American shows:

Does racial diversity in content play a part in viewing preferences? *As an Indian, am I more likely to watch a movie if the thumbnail showed an Indian character in it?*

This became the perfect Final Project idea for our team in the Causal Inference course at the University of Minnesota's MS in Analytics program. **Piyush Gupta, Pushkar Vengurlekar, Olivia Liang, Vibhav Pratap Singh and I** set out on this pipe dream expedition of conducting the perfect randomized experiment.

Experiment Design

Initial version of the experiment we planned out.

We wanted to measure 'interestingness'

Treatment Group: Thumbnails of Movie/Series with someone from their race represented in it.

Control Group: Thumbnail of Movie/Series without someone from their race in it.

People are good at judging something relatively, compared to absolutely.

This was the guiding light when we designed the survey for the recipients. *We did not want them to mark an arbitrary number from 1-10 to say how interesting they thought the movie was.* So we showed them 3 movies, where one embedded image was the Indian/Asian image being tested and asked people to rank the images.

So if the Treatment Group ranked the movies with better racial representation above, my hypothesis would be correct!

So the question became:

Rank the movies in order of interestingness

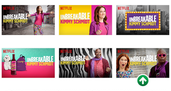
We have two images from the same show, Altered Carbon with an Asian character in the thumbnail, and without one in the Control group

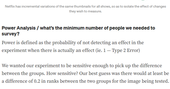
The image from the same show Altered Carbon has been ranked higher in Treatment, which showed an Asian character, compared to Control

Dealing with Confounds

This is where the messiness begins. Unlike a company like Netflix, which has rich information about each of its users, we pretty much didn't have any information. So we wanted to present images of movies/shows that looked comparably similar to each other and were relatively less known to people. If survey respondents had an idea about the show, that would bias our findings.

And unlike Netflix, we did not have a UX team working with us on designing small increments of change with thumbnails so as to isolate the effect of just the changes. Hence, those were the problems we had to acknowledge and work around in our experiment design.





Netflix has incremental variations of the same thumbnails for all shows, so as to isolate the effect of changes they wish to measure.

Power Analysis / what's the minimum number of people we needed to survey?

Power is defined as the probability of not detecting an effect in the experiment when there is actually an effect (in 1 — Type 2 Error)

We wanted our experiment to be sensitive enough to pick up the difference between the groups. How sensitive? Our best guess was there would at least be a difference of 0.2 in ranks between the two groups for the image being tested.

```
> power.t.test(d=0.2,type="two.sample",power=0.8,sig.level=0.05,delta=0.2)

     Two-sample t test power calculation

              n = 393.4765
          delta = 0.2
             sd = 1
      sig.level = 0.05
          power = 0.8
    alternative = two.sided

NOTE: n is number in *each* group
```

To measure a 0.1 difference in ranks with 90% confidence and 80% Power, we needed to collect ~400 samples. The problem is we had about100 students in our class.

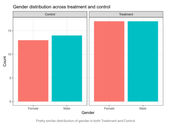
Fix: Ask these 100 students 6 questions each!

People are noisy™

This deals with the 'randomized' part of the randomized-controlled experiment. We needed to have a Treatment and Control group that was completely formed out of random assignment — with no bias towards gender, race, age and other parameters.

The 'other parameters' is the noise that I'm talking about. People are so diverse, especially at a university like the University of Minnesota — so many different locations, with such different life experiences. We decided to interview only our fellow graduate students at the MS Analytics program because we believed it would bring in an element of homogeneity into our groups. But the trade-off was that we would not be able to generalize our insights about the world.

Block Randomization was the approach we used to find our groups — randomized within each factor — Gender, Race, and Age



Pretty similar distribution of gender in both Treatment and Control

Results

Ah, the sweet smell of disappointment

We were not able to measure any differences between the two groups after performing a t-test. In fact, we saw that people were ranking movies with lesser racial diversity lower than the control version. This was absolutely puzzling as all of us expected a treatment effect to exist.

Are we measuring the right KPI?

I started to wonder if people were still arbitrarily ranking movies. After speaking to some of the recipients, I understood they had difficulty giving rank to the movies without any information except the images.

So maybe people aren't able to assign a rank between 1-5, but they are able to decide whether they wanna move a thumbnail up in ranks or not?

We decided to look at each of the groups individually and understand whether they moved the image up or not, very well knowing that this would massively underpower our experiment — meaning we would not be able to support our results with any statistical significance.

The results were nevertheless surprising.

Confirmation biases aside, the fellow Indians in my team agreed strongly with the results, that they would not find movies with Indian characters more interesting.

Yet, all that we could statistically conclude was that we needed to collect more samples to be sure about the direction of these results ie. take it with a **pinch** here! of salt.

So how would we perfect the experiment, given another shot?

1. Handle biases and confounds in images: We have so many factors to control here - image lighting, number of people and so on. The best thing to do would be to work with a UX team and design small increments of change that we can reliably measure.
2. Tired Respondents: Asking 100 people to rank 3 movies six times each is not an easy job. People were reasonably tired. The best fix for this would be to have a larger pool of respondents. Bonus points: We can better generalize about the real world.

3. Survey Design: We used Qualtrics to host our Survey, but there were numerous problems with the ranking interface in the mobile version.

> *The problem with this is people might not be ranking our test image above, because they feel it difficult to do it on the interface!*

Team: Sudharshan Ashok, Piyush Gupta, Pushkar Vengurlekar, Vibhav Pratap Singh, Olivia Liang

More from Sudharshan
Follow
Data Science, Machine Learning, Singer/Music Producer https://sudharshan-ashok.github.io/ https://www.linkedin.com/in/sudharshanashok/

Published in Analytics Vidhya · Feb 23, 2020

Finding the next Billboard #1 — Spotify API Exploratory Analysis
Over past: Analytics and Data Science has been changing the music industry in ways never seen before, through the areas of Music ...
Spotify · 7 min read

Share your ideas with millions of readers. Write on Medium

Mar 12, 2016
Finally here
Plans to blog after long. Maybe I'm done with writing for pay?
2 min read

Love podcasts or audiobooks? Learn on the go with our new app. Try Knowable

Get started
Sign In
Search
Sudharshan
117 Followers
Data Science, Machine Learning, Singer/Music Producer https://sudharshan-ashok.github.io/ https://www.linkedin.com/in/sudharshanashok/
Follow
More from Medium
Lizzy Rain in ILLUMINATION
Elon Musk Eliminated Remote Work Because Working From Home "Doesn't Work"
Bijoshy in Personal or Pixel Eng.
My Salary Increased 12 Times in 5 Years — Here is How I Did It
Peter Lin
AB testing—DID Analysis
Tim Hagan in Learn Dr Comig
Elon Musk Fires 50% Twitter, and Why I Like It
Sudharshan
117 Followers
Follow
More from Medium
Lizzy Rain in ILLUMINATION
Elon Musk Eliminated Remote Work Because Working From Home "Doesn't Work"
Bijoshy in Personal or Pixel Eng.
My Salary Increased 12 Times in 5 Years — Here is How I Did It
Peter Lin
AB testing—DID Analysis
Tim Hagan in Learn Dr Comig
Elon Musk Fires 50% Twitter, and Why I Like It
Help Status Writers Blog Careers Privacy Terms About Text to speech