Question Type: Product

Duration: 40 Minutes
Difficulty: Medium
Domains: Product

Problem

Currently on Netflix, users can watch a movie trailer upon clicking the “preview” button. Netflix wants to improve engagement using an auto-preview. When a user is browsing and a cursor hovers a thumbnail, a preview plays automatically. Should Netflix launch auto-preview?

Solution

> Currently on Netflix, users can watch a movie trailer upon clicking the “preview” button. Netflix wants to improve engagement using an auto-preview. When a user is browsing and a cursor hovers a thumbnail, a preview plays automatically. Should Netflix launch auto-preview?

[Candidate] I’d like to began with questions about the preview. I’m assuming that the preview functions like a movie trailer. When a user hovers a cursor over a video thumbnail, an excerpt is played. Is this correct?

[Interviewer] Yes, that’s correct.

[Candidate] And, in the current browning experience, the user has to click a trailer button to watch the preview, correct?

[Interviewer] That’s right.

[Candidate] I see. It seems to me that the purpose of auto-preview is to help users find videos that they would end up watching.

[Interviewer] Interesting point. Could you elaborate?

[Candidate] Well, If the user plays a movie he’s not sure about, then he might bounce after watching it for just a minute or two. At the sametime, clicking the trailer button might pose some friction for users who want to view snippets quicker.

[Interviewer] That’s great. Can you propose a metric?

[Candidate] Certainly, I think one of the metrics could be the watch time rate which is the total watch time over the browsing time per user session. The idea is that as users spend less time browsing and more time watching, they are engaged more.

[Interviewer] Okay, sound good. Can you design your experiment?

[Candidate] Certainly, I’d first set-up the hypothesis statements. The null hypothesis is that the watch-time-rates of control (manual preview) and variation (auto-preview) groups are the same. The alternative is that the rates differ. I’d calculate the sample size based on alpha = 0.05, power = 0.80 and MDE of 1% lift.

[Interviewer] Before you continue, let me ask a quick question on your choice of MDE. Why 1%?

[Candidate] For a platform like Netflix with +100M MAUs, 1% increase in the engagement metric is practically significant.

[Interviewer] Okay, sounds good. Proceed with how you would set up your experiment.

[Candidate] With the sample size determined, I'd set up the experimentation time (in weeks) based on the traffic I'm willing to allocate to the experiment. Usually a testing of a product feature should take sometime between 1 to 2 weeks.

[Interviewer] What happens if the experiment is less than a week?

[Candidate] Well, there might be variability of user behavior given the day of the week. To remove this confounding effect, it's generally best practice to commit to at least a week.

[Interviewer] Sounds good. What's your statistical test and how would you evaluate it?

[Candidate] The metric is rate which is averaged across users per group. Hence, the appropriate statistical test is the two-sampled T-test for comparing population means. I'd look at the CI of the test. If the direction is an interval in the positive range without zero, then conclude that there is statistical significance at 0.05 that the auto-preview increases the watch-time rate. Finally, either ramp-up with increased power or launch the product.

Interviewer Solution

Should Netflix launch auto-preview? This question involves part-insights and part-AB testing to make a product decision. Every element in a product has a purpose. In the case of auto-preview, the main idea is to facilitate user browsing and help them find movies and shows faster. Ultimately, this would help users stay engaged on Netflix and continue to subscribe - the primary metric of the Netflix's business model.

Insights

Let's start with insights. Before launching the auto-preview or running an AB testing which consumes engineering resource, why should Netflix consider this? Using user log data, you can run the following analysis:

1. What's the average browsing time per user session?
   a. Are users with longer browsing time more likely to churn?
2. What's the bounce rate on videos? (< 10 minutes of watch time on a video)
   a. What percentage of users leave Netflix after watching a video?
3. What's the average watch time per session?

Based on analysis, you can define the following - metric, hypothesis test, experiment, analysis, and product decision.

Metrics

What are you measuring? Defining the metric of an experiment may seem simple, but sometimes it is complex. Choosing the wrong metric produces a vanity value that does not provide a meaningful product decision. With this point in mind, let's construct a meaningful metric.

Let's start with a simple metric: **Watch Time per User Session**.

This metric seems reasonable until you realize its flaws. Before proceeding, can you think of one?

The purpose of auto-preview is to remove friction in helping users find the right movie to watch. Let's consider two users who each spent 60 minutes watching videos. User A flipped through 6 videos, spending just 10 minutes per video before logging out. User B spent the entire 60 minutes on one video. Which user had a better experience? It's clearly user B given that user A was shuffling one video to the next searching for one to watch to no avail. Hence, the watch time per user session is inadequate.

Recall that the purpose of auto-preview is to help users find movies they will enjoy watching. The auto-preview should aim to reduce bounce rate (watch-time less than 5 minutes) and reduce browsing time. Hence, here are two metrics:

1. **Watch Success Rate** = Videos Watched (>= 5 minutes) / Videos Clicked
2. **Watch Time Rate** = Watch Time / (Browsing Time + Watch Time)

The watch success rate measures whether the preview helped users find videos they completed. The watch time rate compares the watch and browsing time. Both metrics increase as users find what they like.

Hypothesis Testing

Given that there are two metrics proposed, two sets of hypothesis statements are required:

Watch Success Rate

Ho: The watch success rate of the control and variation groups are the same.
Ho: The watch success rate of the control and variation groups are different.

Watch Time Rate

Ho: The watch time rate of the control and variation groups are the same.
Ho: The watch time rate of the control and variation groups are different.

The industry standard for significance level is 5% and power being 80-90%. Given that Netflix is a large platform with millions of MAUs, 1% lift is practically significant (MDE).

Experimentation

To run the experiment, the following is required:

1. Randomization Unit - Should you randomize sessions, users, or devices? In this problem, we care about user behaviors so the randomization is at the user level. But, only their first session upon entering the experiment will be measured in the experiment. Why the first session? We will be using T-test which assumes independence of observations. Hence, we will only track the first session of a user.
2. Sample size determination - Given the significance level, power and MDE, you can determine the sample size required to detect an effect in a statistical test.

3. Experiment Duration - The duration of an experiment can be calculated based on the following - total sample size (variation + control) / traffic per week (allocated in the experiment). Suppose that the total sample size required is 50K. 25K users are allocated in the experiment per week, then the total experimentation time is 2 weeks. Generally, the faster the experimentation time better so 1 week is usually ideal.

Analysis

Statistical Test

The main metric is rate which is averaged across users in each group. Therefore, T-test for two-sampled means is an appropriate statistical test.

Given that multiple hypothesis tests are evaluated, a correction method is required to reduce the type 1 error rate. A multiple test correction method such as the Bonferroni Correction is appropriate.

Results

Methods	Sample Size	Metric 1	Metric 1 P-Value	Metric 1 CI	Metric 2	Metric 2 P-Value	Metric 2 CI
Control	25K	0.33	0.0452	(0.005, 0.025)	10.23	0.033	(0.5, 1.7)
Variation (Auto-Preview)	25K	0.35			11.33		

Product Decision

If both metrics increase with statistical signifiance, then launch the feature. If at least one fails then re-design UI or re-run the experiment with increased power.

Interviewer Assessment

In the statistics section, a candidate is assessed based on correctness and soundness of statistical methodology, product sense and communication. For each dimension the candidate is rated in the following scale: (5) superior, (4) good, (3) adequate, (2) marginal, (1) not competent.

Assessments	Rating	Comments
Statistical Methodology	5	The candidate understands the end-to-end process of running an experimentation. She described the steps from hypothesis testing, experimentation, analysis to decisions. When asked follow-up questions on MDE and experimentation time, she had responses that were practically sound. She clearly seems to have experience in running an experimentation which is required for the engagement role.
Product Sense	4	The candidate demonstrated a decent sense in the Netflix platform given that (1) she seemed to understand what would be considered practically significant when asked about MDE and (2) she understood the problem and made correct assumptions about the current browsing experience. One area of improvement is to elaborate more on metrics. There are trade-offs on almost every metric. She chose watch-time-rate as the primary metric. However, this metric alone may not reflect the entirety of the auto-preview experience.
Communication	5	The candidate understood the problem and followed-up with assumptions before proposing a solution. Her thought process was well-explained in a concise manner.