Question Type: Product

Duration: 20 Minutes
Difficulty: Medium
Domains: Product

Problem

Suppose a rideshare platform such as Uber considers offering discounts to riders. The business team hypothesizes that discounts will improve rider engagement on the platform. How would you design an experimentation to test the hypothesis and provide a recommendation?

Solution

[Candidate] I'd like to began with assumptions on three aspects - the discount feature, engagement metric and rider segments. Once I discuss the assumptions, I can propose a methodology.

[Interviewer] Sounds good.

[Candidate] First, I'm going to assume that, for the sake of simplicity, the incentive offered is a one-time 10% discount on non-premium rides. At the moment, the choice of 10% discount is arbitrary, the discount could be 5% or 20%. Perhaps, the next iteration of the experiment could devise an algorithmic approach to optimize the right amount of discount per rider. I'm also focusing on non-premium rides as the bulk of your customer base use the main service similiar to Uber X. This also removes confounding error as behaviors on how riders react to incentives could differ from one class of rides to another.

[Interviewer] Makes sense. What are your thoughts on engagement?

[Candidate] Well, I need to define a proxy for "engagement." Engagement could mean just entering the mobile app without ordering a ride or actually ordering one. In this instance, the engagement is whether the user ordered a ride once a discount is applied.

[Interviewer] Okay. But, on a rider-share app, the user enters the app usually with the intention of ordering a ride regardless of the discount. So, wouldn't your metric merely reflect short-term behavior but, not long-term?

[Candidate] Absolutely. This leads to my next point on users. I think the focus point is to test the discount on, not all, but just first-time users. I assess three possible outcomes among first-time users: (1) a user downloads the app but, never uses the service, (2) a user rides once then churns, (3) a user enjoys the first ride then becomes an active user. The third is the most desired so the metric is the % number of first-time users who rides for the second time. On a commuter app, an active user rides once a week. The time between first-time and second-time must be within 7 days for a user to be considered "active."

[Interviewer] Okay, how would you determine your sample size and experimentation time?

[Candidate] As a general standard, I would choose statistical significance at 0.05 and power 0.8. The minimum effect size is 1%. Based on the number of first-time users entering on a weekly basis, I would allocate % of new visitors for the experiment. Let's say that the initial percentage is somewhere between 5% to 10%. The user will then be randomly assigned to the control (no discount) and variation (10% discount). The experiment duration in weeks is the

sample size * 2 (groups) / number-of-new-users-per-week. I would seek to pick a range that's about two weeks but, no less than 1 week or no more than four weeks.

[Interviewer] Gotcha, how would you conduct your statistical test?

[Candidate] Given that the metric is the proportion of first-time users who rides for the second-time, Two-proportion for T-test is appropriate. The T-test is pooled or un-pooled based on the variance of the two populations. If the p-value is less than the significance level at 0.05, then reject the null hypothesis that there is no difference between the users with or without discounts and conclude that discount works.

[Interviewer] Suppose that in the middle of the experiment, your p-value is less than significance level, can you stop the experiment with the sample conclusion?

[Candidate] Well, if you stop your experiment, you might have achieved statistical significance at that instantaneous point but, it could be a fluke. When sample-size is low, the point estimate, the difference between the sample means, hasn't converged toward the true parameter. Hence, in the beginning of the experiment, the point estimate will vary quite a bit and, the p-value will cross the significance level. But, the conclusion on hypothesis should not happen until the end of the experiment.

[Interviewer] Good point. What would be your final recommendation to the business?

[Candidate] If the test rejects the null hypothesis, I would recommend that discounts should be use to incentivize first-time users to become active users. If the test fails to reject, I would suggest running another experiment with a different promotion offer. Perhaps, instead of 10%, the discount rate could be 20%. Or, instead of applying discount on the first-ride, it could be applied on the second ride to encourage recurrence of ridership.

> **Interviewer Solution**
>
> An ideal response discusses the product problem and experimentation design. The first understand how the discount might incentivize riders. In the short term, they might be incentivized to use the service. However, in the long-term they may not.
>
> Hence, the key metric or the definition of "success" must reflect a long-term behavior. Consider these two metrics:
>
> 1. Percentage of users who requested rides
> 2. Percentage of users who requested rides the second time since the previous rides 7 days ago

What metric best reflects a long-term behavior? It's clearly option 2. This metric measures whether users will become "active" users after receiving an incentive in the form of discounts.

Hence, in general, when designing AB testing, choosing a key metric that serves as a gauge for long-term user behavior and company's business model is vital.

In terms of the AB testing design, the ideal response should flesh out the following:

1. Hypothesis Testing - Ho and Ha statements, Alpha (0.05), Power (0.80)
2. Sample-Size Determination - Effect size (for a large platform such as uber, 1% lift is practically significant)
3. Experimentation Duration - 1 to 2 weeks
4. Randomization - Users
5. Experimental Result - Analysis of lift, p-value, confidence interval
6. Business Rule - If statistically significant, roll out the feature. If fail to reject, then re-iterate with a different feature idea 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 has demonstrated a strong grasp in metric and AB testing designs. He understands that a metric should convey a long-term behavior, not a short-term one. He has also demonstrated that he understands the key steps in AB testing from hypothesis testing, experimentation design and result evaluation.
Product Sense	5	The candidate understands what is considered an active Uber user. The goal for any feature change is to ensure that the users are riding with Uber on a recurrent basis. This is a key point in the business model of Uber. Understanding the premise helped the candidate offer a metric that makes sense - measure whether the discounted ride incentivize the user to ride it the second time without it.
Communication	5	The candidate explicated his steps comprehensive and clearly. He also laid out assumptions in the beginning of the discussion which served as a pivotal steps for his methodology.