Product

Duration: 10 Minutes
Difficulty: Medium
Domain: Product

Problem

A travel app like HotelTonight ran a two-week experiment that showed that a new booking page improved conversion by 33%. However, after launching the new page, the conversion rate flipped, favoring the old version. What could cause this change?

Demonstration

Commentary: The condition for an experimentation is rarely perfect. A variety of reasons during or after an experimentation could skew the primary metric of a launch. Go through the interview below as the candidate addresses one question after another.

[Candidate] When the experimental result fails to emerge after launch, the factors contributing the issue is statistical, technical or external. Which one of those three should I assume?

[Interviewer] Good question. In this interview round, let's focus on possible external factors. Can you elaborate?

[Candidate] Certainly. The timing of launch matters. It is possible that the experimentation took place during the holiday season such as the New Year when visitors are more likely to travel than any other seasons. Such phenomenom is called the holiday effect - a change in a metric because of a holiday. To mitigate the effect, run an experiment on a non-holiday.

[Interviewer] Okay. Can you think of additional external factors?

[Candidate] Certainly, I can think of two more factors. An economic impact because of a recession or pandemic can disrupt consumer spending on travels; thereby, reduce conversions of both old and new page versions. Lastly, a competitor platform might have launched a promotion that moves visitor traffic away from HotelTonight.

Commentary: Notice the breath of external factors that the candidate outlines. A lacking response would have just provided one possible factor or none at all. Holiday effects (or sometimes called seasonal effects), economic disruption (i.e. the 2020 coronavirus pandemic) and competition campaigns are all valid reasons that can disrupt a platform.

[Interviewer] You outlined three factors - holiday effects, economic recession and competition. How would you measure the effect of an external factor on the booking rate?

[Candidate] Well, I see two possible ways to assess external impact on the booking rate. First approach is fitting a regression model: rate ~ time + intervention + time * intervention. The intervention predictor is a dummy variable with pre-external impact and post-external impact. The interaction effect should provide the slopes of before and after the onset of an external factor. I can evaluate the coefficients of the slopes and statistical significance to measure the direction and conclude whether the external impact is significant.

[Interviewer] Good approach. What's your another approach?

[Candidate] The another approach is building a forecasting model such as ARIMA that trains on pre-external impact observations and forecasts on post-external impact period. I could compare the forecast versus the actual using sample-means T-test, probably corrected for autocorrelation, to evaluate if the difference between the expected versus the actual is statistically significant. Mean absolute percentage error between the actual and the forecast provides to what degree the economic impact changed conversion rate overtime.

Commentary: The candidate's response was specific and on-point from the beginning to the end. Note that issues in A/B testing is not strictly statistical and technical errors. Sometimes, external forces cause disruption on a platform. One concrete example is COVID19. When preparing for an analytic and experimentation interview, do not just focus on methodology but, also external factors that can influence key metrics of a platform.

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		
Product Sense		
Communication		

Solution

An experimental outcome failing to generalize is a common pitfall in A/B testing. Various internal and external factors of an experimentation can lead to a false statistical and business conclusion. Let's review the case question:

> A travel app like HotelTonight ran a two-week experiment that showed that a new booking page improved conversion by 33%. However, after launching the new page, the conversion rate flipped, favoring the old version. What could cause this change?

As you read this problem, you should formulate a framework on how you will engage the interviewer. When an experimental result fails to generalize, usually one or multiple validity threats are in play as listed below:

Statistical Errors

1. **Type 1 Error** - The significance level determines the type 1 error rate regardless of the sample size. If the type 1 error rate is 0.05, that's 5% probability of falsely rejecting the null hypothesis and concluding that the alternative is statistically significant. In other words, the experimental effect is due to random chance, not the treatment effect.

   Usually, the significance level in online testing is 5%. It might be possible that the experimenter might have set the significance level too high (i.e. 10%), which also means the higher chance of type 1 error.

2. **Low Sample Size** - If the sample size is low, the power of the test is low. The power is the probability of detecting an effect given that the alternative hypothesis is true. Increasing the sample size improves power. The industry standard is 80-90% power.

   In this problem, having low power is not the cause of experimental effect failing to generalize. Low power of a test means reducing the probability of detecting an effect. However, in this problem, the experimentation produced detected an effect, but failed to carryover. Nonetheless, this is one of the internal validity threats to consider in other A/B testing pitfalls.

3. **Violations of Statistical Test** - If the experimenter utilized T-test, the test assumptions must be checked. T-test assumes independent, identical observations from a normal distribution. The violation of this condition can lead to false rejection of the null hypothesis.

4. **Ending an experimentation prematurely** - It could be possible that the experimenter ended the online testing prematurely upon achieving statistical significance. If 10,000 samples are designated per group to achieve 80% power, the experimentation should collect and measure all the observations, not partial.

Computational Errors

1. **Randomization** - Randomization ensures that there is no sampling bias in an experimentation. If the composition of two cohorts differ, and the treatment favors one type of cohort, then the bias will lead to a false statistical outcome. Diagnostics on the randomization algorithm can reduce the chance of sampling bias.

2. **Performance Issue** - Perhaps the new UI upon launch introduced bugs, not observed in the experimentation. The bugs could explain the reduction in conversion rate. Monitoring performance and logs ensures that the UI/UX of the page in pre-and-post launch remain consistent.

Extraneous Factors

1. **Holiday Effect** - Perhaps the experimentation ran over the course of a holiday (i.e. Christmas and New Years Eve), which bolstered the effect of treatment on the variation group. Unless the experimentation is designed to deliberately test an effect during the holidays, it's usually a bad idea as the result will fail to generalize on most other days of a year.

2. **Seasonality Effect** - Perhaps, the experimentation only ran over the weekend. The day of the week may have an impact on the experimentation. It's usually ideal to run at least 7 days, and at most 14 days to reduce the seasonality effect.

3. **Novelty Effect** - The novelty effect is a psychological phenomenon when a user reacts positively to a change. However, the effect is not lasting, just temporary. To mitigate the novelty effect, longer experimentation time or sampling new users is an approach.

4. **Competition** - It could be possible that a competitor event (i.e. promotions) could temporarily cause the treatment effect to magnify or diminish during an experimentation. Running an experimentation when there's no competitor event is vital.