Section 4 - Experimental Design

The original version of an ecommerce website is quite basic, featuring only text on the site. They plan to add some product images, and intend to run an A/B test on their website to see if it helps sales.

a. Hypothesis –

Adding product images will increase conversion rate by *x%*. (x is calculated based on what is the minimum increase in conversion (lift) which will have a sufficient business impact)

b. The primary metric for measuring success would be conversion rate. Conversion rate definition will mainly be conversion from product page views to add-to-cart step (next step in conversion funnel). We can track the top-of-funnel conversion leading to purchase (sales conversion) however this effect may be smaller since page conversion to sales conversion may have many steps which may be causing drops and not necessarily related to our experiment goal .

c. Secondary metrics to consider would be –

- "product page views" – there should be increase in product page views (lift in page views from a before-vs-after analysis for same product as well as comparing page views across similar products of same category which have vs don't have images)
- "add-to-cart (or other CTA) actions" – product images should drive up the propensity to purchase or add-to-cart clicks
- "time spent on product page" – customers would spend more time viewing the product images now
- "bounce-rate on product pages" – should go down as customers will hopefully have more propensity to travel down the conversion funnel
- Page load time/ page load errors – adding product images will increase page size & in turn may affect page load time. These may cause page load errors and increase page reload events and page bounce rate. These should act as guardrails against any negative user experience due to the engineering changes for this experiment
- "average order value (AOV)" – though the overall sales increase, it may reduce the average order value as the products with images may cannibalize customer's attention from similar products which don't have images and these products may be priced differently.

d. Having multiple layers of filters on top of one another may make it increasingly difficult to isolate the effect of the particular experiment to check for causation. Also, with multiple filters, it becomes difficult to ensure user actions get attributed to correct group (treatment vs control) across layers of breakdowns.

e.

Cohort	Visitors	Converters	Traffic split
Variant 1	8000	3000	10%
Control	72000	25000	90%

I would recommend to launch/ adopt the Variant1 as it has a significant improvement over control. I have shared the calculations below.

The probabilities for the two groups are –

$$Pv = \frac{\#converters\ (variant1)}{\#\ visitors\ (variant1)} = \frac{3000}{8000} = 0.375$$

$$Pc = \frac{\#converters\ (control)}{\#\ visitors\ (control)} = \frac{25000}{72000} = 0.347$$

Since we are testing for difference in these probabilities, we need pooled probability of success and pooled variance. So formula becomes –

$$Ppooled = \frac{\#converters\ (variant1) + \#converters\ (control)}{\#\ visitors\ (variant1) + \#\ visitors\ (control)} = \frac{3000 + 25000}{8000 + 72000} = 0.35$$

$$S^2 = Ppooled\ (1 - Ppooled) * (\frac{1}{Nc} + \frac{1}{Nv}) = 0.0000316$$

$$Standard\ Error = \sqrt{S}^2 = 0.0056$$

The test statistics of the 2-sample Z-test for the difference in proportions can be calculated as follows:

$$T = \frac{Pc - Pv}{SE} = -5$$

abs(T) = 5

For 2-tailed z test with significance level **α = 0.05, $Z_{critical}$ = 1.960, p-value = 0.000001**

Since abs(T) > $Z_{critical}$, Variant1 has better performance than control. Hence we can adopt Variant1.

f. It may not make much sense in spending the time & effort in conducting an A/B test in following scenarios –

- If there are not enough users in the required segment/ sufficient traffic to conduct an experiment and successfully measure the effect.

- If the new feature or change is independent and low-risk – in this case the cost of running an experiment might not be justifiable.

- If the feature/change is high priority and there is not enough time to run an experiment

- If the hypothesis to conduct an experiment is not clear

Generally, A/B testing works best to test alternate options of same feature, upgrades to existing features or soft launches of small features. They may not be a great option for product-roadmap changes, product-market fit tests (user research etc. may be more suited) or complete product rehaul. This is because such large changes or completely new, creative out of box solutions may not find meaningful traction within the time and customer segment boundaries of an A/B test.

In case a new feature is directly rolled out, we can do phased roll out to control risk. We would keep an eye on guardrail i.e. product performance metrics, churn and top level primary metrics while continuously monitoring the adoption and engagement for new feature to understand if adoption over a period of time justifies the cost of maintaining the feature.