Bill Su

Final Project A/B Testing

Experiment Design

Metric Choice

I have chosen Number of Cookies and Click-Through Probability as the invariant metrics of this study, and Gross Conversion and Net Conversion as Evaluation Metrics.

Number of cookies is a good invariant metric because it does not change regardless of the outcome of the experiment as the experiment condition are applied after cookies are initiated. Being a good invariant making it a bad evaluation metric.

Number of user ids enrolled in the course will be different between control and exp group since the experimental condition was applied before the user signs up for the free trials. Therefore, it cannot be an invariant metric. It could potentially be an evaluation metric, but conversion rate is much better as evaluation metrics than flat number since they provide more information.

Number of clicks could potentially be another invariant metric, but it essentially provides the same information as click through probability and click through probability provide a probability, which is a better metric in my view. At the same time, having too much invariant metrics might result in higher chances of false positives during the sanity check, so I did not include it. Being a decent invariant metric making it a bad evaluation metric.

Click through probability is a good invariant metric because it occurred before the conditions were applied and can diagnose engineering errors when changing the click buttons. Being a good invariant making it a bad evaluation metric.

Gross conversion is our first evaluation metric. It identifies whether the percent of free trail enrollment decrease as a result of the changes. Obviously it is a bad invariant due to being a good variant. I expect this value to go down significantly as a result of the change because users might be deterred by the time commitment message. **This metric is expected to go down since the message deterred users away.**

Retention might be a good evaluation metric, identifying the increase/decrease of checkout rate after users signing up for the free trial. However, retention rate can be calculated or inferred from both gross and net conversion. For example, if gross conversion goes down and net conversion remain the same, we can expect retention to go up. Therefore, to prevent having too many metrics, I decided to not include it as an evaluation metric.

Net conversion is our second evaluation metric. It provides new information in addition to Gross conversion regarding whether users ended up going to checkout more as result of our changes. Again, bad invariant. **I expect this value to do up as result of better user experiences during trials.**

Bottom line, if net conversion goes down, the experiment should not be launched. On the other side, if gross conversion goes down and net conversion does not go down, the experiment should defiantly be launched. Everything in the middle require more detailed follow-up.

Measuring Standard Deviation

Gross Conversion: .0202
Net Conversion: .0156

The unit of diversion, which are cookies, is the same as the unit of analysis (number of unique cookies clicking the button) for both metrics. Therefore, we should be able to establish similarity between empirical and analytical variance.

Sizing

Number of Samples vs. Power

I will not use Bonferroni correction during my analysis because I am only using two metrics. I will need approximately 685,325 pageviews in my experiment.

Duration vs. Exposure

I do not want to divert too much traffic from Udacity traffic since the experiment might result in significantly lower Gross and Net Conversion, which harms Udacity's user experience and revenue within the duration of the experiment. If we were to use only 60% of the traffic, only 30% of the users will be exposed to the experiment group with 29 days of experiment time. The long duration of this experiment will enable rapid evaluation and termination if the change drove down too much revenue. Therefore, the risk of this experiment in terms of revenue lost is minimal, and there are no significant other risks.

Experiment Analysis

Sanity Checks

Number of Cookies: [.49882, .50118], observed: .50064, passes: YES!
Click Through Probability:[.49588, .50411], observed: .50047, passes: YES!

Result Analysis

Effect Size Tests

Gross Conversion: [-.02912, -.01199], statistical significance: YES!, practical significance: YES!
Net Conversion:[-.01160, .00186], statistical significant: NO ☹, practical significance: NO ☹

Sign Tests

Gross Conversion: p = .0026, statistical significance: YES!
Net Conversion: p = .6776, statistical significance: NO ☹

Summary

I did not use Bonferroni correction because this is an All/And situation instead of an Any/Or situation. Bonferroni prevents the situation in which the rejection or the significance of ANY metrics will result in the incorrect launch of the experiment, but in this case both metrics need to be satisfied to launch the experiment, effectively making Bonferroni unnecessary.

There are no discrepancies between the effect size and my sign tests.

Recommendation

The experiment condition in the A/B test decreased gross conversion rate while did not change the net conversion rate. However, as we observe the CI for net conversions, we have realized that it is dangerously close to the negative boarder of significance. A negative net conversion will decrease Udacity's revenue collected from users and is a major setback. Therefore, I despite the fact of achieving desired result, I would not recommend launching the experiment. Instead, Udacity should launch another similar experiment with more sample size and more power to determine whether net conversion is really going down.

Follow-Up Experiment

Another way of decreasing early cancellation is to send follow up emails to users who have not logged into Udacity for classes during free trial period for at least 2 days and offer them a 2-day extension of their free trials. This will encourage the free trial users to become active learners again and eventually convert. The unit of diversion should be unique user IDs in this case because the unit of analysis of our study is user as well. The invariant metric could be unique user IDs during the duration of the experiment and the evaluation metric could be retention rate after free trial, calculated by users who did not cancel free trial/ total users signed up for the free trial. My hypothesis is that the free trial extension will significantly raise retention rate of users during free trial.