P/7 Design an A/B Test

Edmund Wong

Experiment Design

Metric Choice

List which metrics you will use as invariant metrics and evaluation metrics here.

- **Invariant metrics:** Number of cookies, Number of clicks, Click-through-probability,
  Evaluation metrics: Gross conversion, Retention, Net conversion

For each metric, explain both why you did or did not use it as an invariant metric and why you did or did not use it as an evaluation metric. Also, state what results you will look for in your evaluation metrics in order to launch the experiment.

- Because the screener pops-up after clicking on 'start free trial', number of cookies, number of clicks, and click-through-probability were used as invariant metrics. Since all the other metrics could be affected after user interacts with the screener, those metrics were not used as invariant metrics.

Gross conversion, retention, and net conversion were used as evaluation metrics because these are all rate metrics that involve post-interaction with the screener and are expected to change between the control and the experiment groups. User-id is also expected to change but might not be a good evaluation metric because it is a count rather than a rate.

A significant decrease in gross conversion supports part of the hypothesis, that the number of students that are frustrated in the free trial will be reduced. A significant increase in retention also supports the hypothesis that there will be a smaller percentage of enrolled students who will leave the trial. A significant increase in net conversion supports part of the hypothesis that the experiment will not significantly reduce the number of students who will continue past the free trial.

Measuring Standard Deviation

List the standard deviation of each of your evaluation metrics.

- **Gross conversion SD:** 0.0202
- **Retention SD:** 0.0549
- **Net conversion SD:** 0.0156

For each of your evaluation metrics, indicate whether you think the analytic estimate would be comparable to the empirical variability, or whether you expect them to be different (in which *case it might be worth doing an empirical estimate if there is time). Briefly give your reasoning in each case.*

Since our unit of diversion in this experiment is cookies, gross conversion and net conversion (in which their unit of diversion is also cookies) are metrics that the analytic estimate would be comparable to the empirical variability. Retention, which has a unit of diversion in user-id, is expected to be different.

Sizing

Number of Samples vs. Power

Indicate whether you will use the Bonferroni correction during your analysis phase, and give the number of pageviews you will need to power you experiment appropriately.

Bonferroni correction was not used for the analysis phase.

Retention will require 4741213 pageviews, but including this metric in the experiment will take too long to run, as the number of pageviews per day, 40000, is too small relative to it. Therefore, I will not analyze this experiment using retention. Because the evaluation metrics selected were gross conversion and net conversion, 685325 pageviews will be needed to power the experiment appropriately.

Duration vs. Exposure

Indicate what fraction of traffic you would divert to this experiment and, given this, how many days you would need to run the experiment.

I would divert 75% of traffic to this experiment, using 23 days to run the experiment.

Give your reasoning for the fraction you chose to divert. How risky do you think this experiment would be for Udacity?

I believe that this experiment is not that risky for Udacity, since the participants merely just need to answer one extra question (which isn't even considered sensitive). Because of this low risk, I believe 75% diversion is safe without being too risky.

Experiment Analysis

Sanity Checks

For each of your invariant metrics, give the 95% confidence interval for the value you expect to observe, the actual observed value, and whether the metric passes your sanity check.

- Number of cookies (LB: 0.4988, UB: 0.5012, Observed: 0.5006)
- Number of clicks on "Start free trial" (LB: 0.4959, UB: 0.5041, Observed: 0.5005)
- Click-through-probability (LB: 0.0812, UB: 0.0830, Observed: 0.0822)

All metrics passed because the observed values are within each of the confidence intervals.

Result Analysis

Effect Size Tests

For each of your evaluation metrics, give a 95% confidence interval around the difference between the experiment and control groups. Indicate whether each metric is statistically and practically significant.

- Gross conversion (LB: -0.0291, UB: -0.0120) Yes it is both statistically and practically significant
- Net conversion (LB: -0.0116, UB: 0.0019) No it is neither statistically and practically signficant

Sign Tests

For each of your evaluation metrics, do a sign test using the day-by-day data, and report the p-value of the sign test and whether the result is statistically significant.

- Gross conversion p-value: 0.0026
  $p < 0.05$, therefore statistically significant
- Net conversion p-value: 0.6776
  $p > 0.05$, therefore not statistically significant

Summary

State whether you used the Bonferroni correction, and explain why or why not. If there are any discrepancies between the effect size hypothesis tests and the sign tests, describe the discrepancy and why you think it arose.

Bonferroni correction was not used because we are relying on all of our metrics to be significant in order to support the hypothesis. For the evaluation metrics, the effect size tests have shown that gross conversion, the number of enrolls divided by the number of cookies, were shown to be both statistically and practically significant. In contrast net conversion, the number of users that remained past the 14-day mark divided by the number of cookies, were shown to be neither statistically and practically significant. The sign tests have shown that gross conversion was statistically significant while net conversion was not. The results of the sign tests did not show any discrepancy with the effect size tests. Both tests agree for both evaluation metrics.

Recommendation

Since there was a significant and practical significance in the decrease of gross conversion, the experiment was successful in reducing the number of frustrated students who left the free trial. In addition, there was no statistical significance in net conversion, but the confidence internal overlaps into the negative side of the practical significance boundary. This means that the number of non-trial enrolled students is possibly reduced and will affect the business. Therefore, it is recommended that this experiment should not be conducted.

Follow-Up Experiment

Give a high-level description of the follow up experiment you would run, what your hypothesis would be, what metrics you would want to measure, what your unit of diversion would be, and your reasoning for these choices.

A follow-up experiment is to add videos during the 14-day period of real stories from candidates who succeeded in the Udacity program and obtained desirable jobs as a result of this program. The hypothesis of this experiment is that these videos will be a motivating factor to encourage students to continue past the 14-day period. Retention is the metric that we will want to measure, which is the number of user-ids to remain enrolled past the 14-day boundary divided by the total number of enrollments. The unit of diversion will be user-ids. All the other possible evaluation metric choices have unit of diversion in clicks, which would not be the ideal choice of measure for this experiment. If retention rate is statistically and practically significant, then it is a good idea to implement the experiment. We can also include number of user-ids as an invariant metric to this experiment, as user-ids is the unit of diversion in this experiment.