A/B Test Design for Free Trial Screener

Matt Ignal
9/26/16

Overview

This experiment will test the efficacy of a screener on Udacity's course homepage. If a student were to click "start free trial", they would be presented with a question: How many hours per week would they devote to the course? If the student entered fewer than five hours, a message would appear that successful completion of the course would require a greater time commitment with a suggestion to instead access the free course materials while giving the option to continue enrollment in the free trial.

If the expectations for time commitment were communicated to students at the outset, there might be greater user satisfaction lead to a greater number of students who continued with the course past the free trial.

The screener can be viewed here:
https://drive.google.com/file/d/0ByAfiG8HpNUMakVrS0s4cGN2TjQ/view

Experiment Design

Metric Choice

- **Number of cookies:** ***Invariant Metric***

  The number of unique cookies to view the course overview page is a good invariant metric for the experiment. We should not expect unique traffic directed to the course overview page to be impacted by the screener since the screener is only displayed after "Start free trial" is clicked on the page. Unique cookies will be split up into equally-sized control and experiment groups to allow for proper comparison of changes.

- **Number of user IDs**: ***Neither Invariant nor Evaluation Metric***

  The number of users who enroll in the free trial should be affected by the implementation of the free trial screener. Udacity asks students to reassess their decision, so we would probably expect fewer students to enroll in the free trial. For this reason, the number of user IDs cannot be invariant. Furthermore, this metric cannot be an evaluation metric because there is not enough useful information in just the number of user IDs to enroll in order to draw conclusions about the impact of the screener. The rate of students who enroll in the course would provide more information.

- **Number of clicks:** ***Invariant Metric***

The number of unique cookies to click the “start free trial” button occurs before the screener, so like the number of cookies, it should be an invariant metric because we would not expect it to be impacted by the implementation of the screener.

- **Click-through-probability: *Invariant Metric***

The probability of unique cookies to click “Start free trial” out of those who view the course overview page is an invariant metric as it incorporates the number of cookies and the number of clicks metric, which are both invariant as they occur before the screener appears. We should not expect the click-through-probability to change.

- **Gross conversion: *Evaluation Metric***

The number of students who enroll in the course divided by the number of unique cookies to click “Start free trial” should be influenced by the experiment. Users are asked to reassess their decision to enroll in the free trial, so there may be a smaller percentage of students to enroll. Gross conversion allows for the determination of the impact of the screener because it demonstrates the difference in the rate of students to click through to the free trial.

- **Retention: *Evaluation Metric***

The amount of user IDs to remain enrolled past the 14-day boundary and thus make at least one payment divided by the amount of user IDs to enroll in the course should be impacted by the implementation of the screener. Measuring this change will allow for evaluation of the effect of the screener on retaining students past the free trial period.

- **Net conversion: *Evaluation Metric***

The number of user IDs to make at least one payment divided by the number of unique cookies to click the “Start free trial” button is arguably the main objective of the implementation of the screener: to improve the user’s experience by screening out less committed students while not incurring a significant reduction in the number of users who pay. The effect of the screener on students remaining in the program is an evaluation metric.

The results for evaluation metrics should answer if the overall student experience is improved by the introduction of the screener by providing them with the expectations for the course up-front without reducing the number of students who pay. The gross conversion metric should then measure the change caused by filtrating students, so we expect it to decrease upon implementation. The retention metric should measure the effect of this filtration by measuring the rate of those who pay.

In order to launch the screener, we need all three results to match our expectations. Gross conversion should decrease because the screener would ideally filter out students who cannot

commit enough hours to Udacity's courses. Retention would then need to rise in order for there to be no significant decrease in the net conversion, as net conversion is the product of gross conversion and retention.

Measuring Standard Deviation

Given the following baseline values (adjusted for 5000 unique cookies to view the page per day):

Unique cookies to view page per day:	5000
Unique cookies to click "Start free trial" per day:	400
Enrollments per day:	82.5
Click-through-probability on "Start free trial":	0.08
Probability of enrolling, given click:	0.20625
Probability of payment, given enroll:	0.53
Probability of payment, given click	0.1093125

The standard deviation of the evaluation metrics can then be calculated for the binomial distribution. This yields:

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

As the metrics are probabilities with binomial distribution with a sample size of 5000, the analytic estimate should be fairly accurate. For gross conversion and net conversion, the metrics' denominators are cookies, and given that cookies are what are being diverted in this experiment, the empirical estimates should be close to the analytical ones. Not so for retention, where the metric's denominator are user IDs. Given that this introduces a little more variability, if there is time, we should use the empirical estimate.

Sizing

Number of Samples vs. Power

Bonferroni correction is not used (explanation can be found in the summary).

(Calculations were completed at http://www.evanmiller.org/ab-testing/sample-size.html using alpha = 0.05 and beta = 0.2)

- **Gross Conversion**: Given a baseline probability of 0.20625, a dmin of 0.01, the minimum sample size was calculated to be 25,835 clicks. Dividing that by the click

through probability (0.08) and multiplying by 2 to account for the two groups, the minimum amount of pageviews is 645,875.

- **Retention**: Given a baseline probability of 0.53, a dmin of 0.01, the minimum sample size was calculated to be 39,115 enrollments. Dividing by the click through probability and then the enrollment probability (0.20625) and again multiplying by 2 produces a minimum amount of pageviews at 4,741,212. This will prove to be too time consuming, *so retention will be dropped as an evaluation metric*.
- **Net Conversion**: Given the baseline probability of 0.1093125, a dmin of 0.0075, the minimum sample size was calculated to be 27,413 clicks. Dividing that by the click through probability and multiplying by 2 produces a minimum amount of pageviews of 685,325. *As this is larger than the minimum of amount of pageviews in gross conversion, this will be the number of pageviews needed.*

Final Evaluation Metrics:

- Gross Conversion
- Net Conversion

Number of Pageviews: 685325

Duration vs. Exposure
Fraction of Traffic Exposed: 1.0
Duration of Experiment: 18 days

685325 divided by 40000 unique cookies per day produces an 18-day experiment duration. There is no good reason why 100% of the traffic cannot be diverted to the experiment as there is very little risk in implementing a simple screener, and at just 2 and ½ weeks, the experiment duration is reasonable given the proposed changes.

Experiment Analysis

Sanity Checks

Computing the total amount of controlled and experimental cookies allows for application of the following formulas:

For a 95% confidence interval, z* is 1.96. m, or the margin of error, gets added or subtracted to the theoretical even split of 0.5 between the number of controlled vs. experiment in the theoretical model.

Click-through probability is 0.08. Since the hypothesis test is on the mean difference, we should use a pooled standard deviation to calculate the change in the margin of error to be applied to the click-through probability:

Using the experimental data with a 95% confidence interval:

- **Number of cookies**: 0.4988 (lower bound), 0.5012 (upper bound), 0.5006 (observed) → **Pass**
- **Number of clicks**: 0.4959 (lower), 0.5041 (upper), 0.5005 (observed) → **Pass**
- **Click-through-probability**: 0.0812 (lower), 0.0830 (upper), 0.0821 (observed) → **Pass**

As the observed values all lie within their respective confidence intervals, the sanity check passes for the invariant metrics.

Result Analysis

Bonferroni correction is not used (explanation can be found in the summary).

Effect Size Tests

As with the sanity checks, pooled standard error was calculated to measure a 95% confidence interval in the difference between the control and experimental groups. Statistical significance would not include zero in the interval, while practical significance would not contain the given dmin (±0.01 for gross conversion, ±0.0075 for net conversion).

Using the experimental data with a 95% confidence interval:

- **Gross Conversion**: -0.0291 (lower bound), -0.0120 (upper bound) → **Statistically and Practically Significant**
- **Net Conversion**: -0.0116 (lower), 0.0019 (upper) → **Statistically and Practically NOT Significant**

Sign Tests

For the 23 days where there was both enrollment and paid data available, the gross and net conversion ratios could be calculated for both control and experimental groups. To be a "success", experiment conversion rates should drop as expected. For 19 out of the 23 days, the conversion rate of the control group was larger than that of the experiment group. For net conversion, the experiment group had a larger conversion rate 10 out of the 23 days (we are expecting this to remain about the same). Applying these values to two-tailed p-value calculations with a hypothetical probability of success of 0.5 yields:

(p-value calculations for sign test were completed at http://graphpad.com/quickcalcs/binomial1.cfm)

- **Gross Conversion**: p-value: 0.0026 → **Statistically Significant**
- **Net Conversion**: p-value: 0.6776 → **Statistically NOT Significant**

The p-value of the gross conversion is 0.0026 is less than alpha (0.05), while the net conversion (0.6776) is greater and is therefore statistically not significant.

Summary

The Bonferroni correction was not used since both evaluation metrics were needed to match expectations in order to launch. While Bonferroni corrections are designed to reduce the risk of a Type-I error (or a false positive), it is ill-suited for Type-II errors (or false negatives). Failing to reject the null hypothesis while multiple metrics are being tested in conjunction will hamper us from making conclusions about whether to launch the screener, so the Bonferroni correction is too conservative for this experiment.

The effect size hypothesis tests and the sign tests showed no discrepancies. While gross conversion was shown to be statistically and practically significant, net conversion was neither.

Recommendation

I would not recommend launching the proposed change until further testing is completed. The screener, which asks users to reassess whether or not they should go ahead with the free trial if they are not able to commit enough hours to programs, failed the A/B test. While gross conversion clearly declined as expected, it could not be established that there was no significant decline in net conversion. The 95% confidence interval for net conversion included the practical significance boundary of (-0.0075), so there is a realistic chance that going ahead with the proposed change might fail to meet Udacity's expectations and could result in a loss of revenue. On the other hand, the large confidence interval suggests that it is still possible for there to be virtually no change in net conversion. Running the experiment for double the duration would likely provide more insight into the amount of change in net conversion the screener causes.

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.

In order to improve user experience, Udacity should continue to investigate ways to improve the retention rate in its Nanodegree programs without negatively impacting net conversion. One way of doing this would be to overhaul the first project that Udacity offers in each of its programs so that students would be more likely to continue. First impressions are important, and I suspect that many students could be turned away for some of the following reasons:

Course is too fast/slow

- Quantifiable progress takes too long
- Learned information can be hard to access

Hypothesis: If Udacity were to offer a brisk first course and project that would quickly demonstrate the value of working through a Nanodegree, then students would be more likely to continue with the program. A series of 3-4 mini-projects as part of the greater course project, similar to those offered in the Data Analyst Nanodegree Intro to Machine Learning course, would best accomplish this goal. Students could work at their own pace, see clear progression

quickly, and refer back to prior work while completing the project. Most importantly, they would "learn by doing."

Metrics:

- Number of user IDs - Invariant
- Number of students to make at least one payment - Neither invariant nor evaluation metric
- Retention - Evaluation metric

We wouldn't expect the number of students to enroll in the trial to be any different because the course content only gets introduced after the trial begins, but we do want the number of students to make a payment to rise. The ratio of number of students with a payment to the number of students to enroll in the trial, or the retention rate, should rise if the proposed change is to be launched.

Unit of Diversion: User IDs should be the unit of diversion since they are the individual subject of the experiment.