Will Roberts

Udacity A/B Test Experiment Design

Metric Choices

Invariants

Number of cookies: Number of unique cookies to view the course overview page. (d_{min}=3000)

This is a great population sizing invariant for the test. The cookies will be randomized between the test and control groups as a result of being the unit of diversion. This group includes the free access program and free trial nanodegree students because it is the first page seen after choosing either of the options.

Number of clicks: Number of unique cookies to click the "Start free trial" button (which happens before the free trial screener is trigger). (d_{min}=240)

Students will click the "Start free trial" button before the control and experiment groups experience different environments. Since both groups must pass through this step for the experiment to trigger, this metric is an invariant metric.

Click-through-probability: Number of unique cookies to click the "Start free trial" button divided by number of unique cookies to view the course overview page. (d_{min}=0.01)

Like the number of clicks above, the click-through-probability here is a metric looking at an event that happens before the screener is triggered. There shouldn't be a difference between the two groups.

Variants

Gross conversion: Number of user-ids to complete checkout and enroll in the free trial divided by number of unique cookies to click the "Start free trial" button. (d_{min}= 0.01)

The experiment hopes to give the student more information before processing payment and fully enrolling in the free trial. The gross conversion is a variant because we are looking to see if a student with more knowledge of what the program entails will be more likely to start with the free programs than full immersion into the nanodegree.

Net conversion: Number of user-ids to remain enrolled past the 14-day boundary (and thus make at least one payment) divided by the number of unique cookies to click the "Start free trial" button. (d_{min}= 0.0075)

The net conversion is a variant metric as well because it represents the student's monetary commitment to the program. The company aims to keep as many paid students in the program as possible under a logical business model, so Udacity would want to ensure that they are not losing a significant number of students in the nanodegrees because of the new screener.

Neither

Number of user-ids: Number of users who enroll in the free trial.
This was not chosen as an invariant because it is assumed the control and experiment groups will be different here. People in the control who click "Start free trial" will be prompted only to the process of creating a log-in. This differs from the experiment group, because these students will have the option of the registration process or asked about taking another program for free depending on time constraints. It was also not chosen as a variant because we are already using gross conversion to measure the impact on enrollments. It's ideal to use that ratio rather than the user-id raw number because it can normalize itself to compare between the experiment and control groups with different numbers of users.

Retention: Number of user-ids to remain enrolled past the 14-day boundary (and thus make at least one payment) divided by number of user-ids to complete checkout.
Retention was not chosen as a variant or invariant because of the time it would take to get enough pageviews to satisfy the metric. It would take 4,741,213 pageviews which equates to 119 days. Though the metric would help determine an impact on students sticking with the program, the time frame is not reasonable for this study.

First off, it must be noted that this test is done with the students' best interests in mind. In a purely monetary sense, it would make sense to sign up as many as possible to the free trial with the hopes of converting them to full time payment. However, this test aims to decrease student frustrations with the program and have a more manageable pool of committed students for the coaches to devote their attention to.

Explaining expectations better in the beginning, the company would experience fewer less committed students in the free trial period. This would help lighten the load for coaches who could focus their attention on students who are paid or more likely to continue with the paid version. In a perfect world, the net conversion would increase resulting in more revenue for the company, but it can't be expected that with fewer people signing up for the free trial that more people will end up staying after the free period. Therefore, it is the hope that the test does not result in a significant decrease in paying customers.

In all, to launch the changes on a full scale, there would need to be a significant negative change in the gross conversion metric (d_{min}= 0.01). Also, there would be no significant negative change to the net conversion (d_{min}= 0.0075).

Measuring Standard Deviations

Gross Conversion: 0.0202
Net Conversion: 0.0156

The Bernoulli distribution with the standard deviation formula sd = sqrt(p * (1-p) / N) was used to obtain the standard deviations above. For each metric, the unit of diversion is cookies. The unit of analysis for gross conversion and net conversion is also cookies. Since these units match, the analytic and empirical variability should be fairly similar.

Sizing

Number of Samples vs. Power

After careful consideration, the Bonferroni correction will not be applied in the analysis. Because we need both the gross conversion to be statistically significant and net conversion to be insignificant, this experiment sits under the umbrella of ALL metrics needing to fall in the specified ranges to consider a launch of the new settings.

Since net conversion requires more pageviews to run the experiment, that number will be used as the minimum pageviews needed (685,325).

Duration vs. Exposure

Since the experiment can be categorized as minimal risk, we confidently can divert all traffic. To justify this, no student is in danger of harm in the program. Also, the study does not collect any extra sensitive information that the company is not already privy to when a student enrolls. Another factor to consider is a possible reduction in revenue while running the experiment since we anticipate the potential for less students to sign up for the nanodegrees. In light of this, it’s imperative to keep the duration of the study as short as we can justify while still collecting everything necessary. Given the necessary pageviews needed (685,325) and page views per day (40,000), the experiment would be run over an 18-day period.

Experiment Analysis

Sanity Checks

The three variants with 95% confidence intervals, observed values, and metric passing are as follows:

Number of Cookies: (0.4988, 0.5012), Observed: 0.5006, PASSED

Number of Clicks: (0.4959, 0.5041), Observed: 0.5005, PASSED

Click-through-probability: (0.0812, 0.0830), Observed: 0.0821, PASSED

Result Analysis

Effect Size Tests

Gross Conversion: (-0.0291, -0.0120)
Because our confidence interval for difference does not include 0, it is statistically significant at a 95% confidence level.
The observed difference in the experiment (-0.0206) is larger than the absolute value of the set minimum (0.01), so the results are practically significant.

Net Conversion: (-0.0116, 0.0019)
The difference is not statistically significant because in this case the confidence interval does contain 0.
The observed difference (-0.0049) is smaller than the low end of the practical difference interval (-0.0075), so the results are not practically significant. However, the low end of the observed confidence interval (-0.0116) exceeds -0.0075. This revelation will be explored further when making a recommendation.

Sign Tests

In the sign test, only 4 out of the 23 days showed a lower gross conversion rate in the control group than the experiment group. The p-value was 0.0026 and statistically significant.

For the net conversions, 10 out of 23 days showed a lower gross conversion rate in the control group than the experiment group. The p-value was 0.6776 and not statistically significant.

Summary

Again, it should be noted that the Bonferroni correction was not applied in the analysis. Our test was searching for a statistically significant decrease in one metric (gross) and no statistically significant decrease in the other (net), and the changes would be considered for full scale launch only if both conditions were met. The correction is too conservative in this case where ALL metrics need to be met because it increases the chances of a false negative, leading to a recommendation to not launch a potentially valid change.

In review, both the effects size test and sign test agree that there was a statistically significant decrease between the control and experiment groups for gross conversion. They also found no significant decrease in the net conversion. With these two results in mind, we reject the null hypothesis.

Recommendation

The hypothesis stated that the screener page following the ‘Start free trial’ button click would set clearer expectations for students, thus reducing the volume of students leaving the free trial without significantly reducing the number of students to continue past the free trial. Our metric gross conversion showed a statistically significant decrease in students enrolling in the free trial after seeing the screener, and the net conversion metric observed no statistically significant decrease in students continuing past the trial period. However, the results give us pause when looking at the confidence interval for net conversion. This metric impacts the company’s bottom line, and it would be hard to recommend a change where a large majority of the interval lies in the negative range. In other words, there is some strong evidence that a launch would bring about a decrease in students who continue to the paid portion of the program. Additionally, the low end of the observed confidence interval extends past the low end of the practical significance interval which insinuates the change could potentially have a negative practical significance impact. With this in mind, an immediate launch is not recommended. The screener did show promise, so more tests are necessary to support a launch.

Follow-Up Experiment

Keeping with the same theme as this study, it is in Udacity’s best interest as a leader in academia to provide students with the best chance for success and longevity in the program. That was one of the main goals here and in future follow-up studies should be always considered near the top of all priorities.

Many companies have found success and created brand loyalty using rewards programs. Therefore, the follow-up study would implement a similar program to encourage students to continue in the paid program and progress through the course. Many students in the Udacity-style learning environment are self-motivated, but this type of positive reinforcement could help incentivize their learning even more. Rewards could range from videos giving insight into the day to day life at many hiring partners to small gifts like music subscriptions or books. Though this would take a little brainstorming and possibly a small financial investment by the company, more long term subscriptions and greater feeling of accomplishment for students would be a win-win for everyone.

The hypothesis is if the company implements a rewards program that recognizes accomplishments and progress of students at preset checkpoints throughout the program, students are more likely to stick with the subscription through frustration in the early stages attempting to unlock rewards. The evaluation metrics would be retention, retention after 3 months (payments), and total students who make at least 1 payment. Retention (same definition as the current study) measures short term engagement in the program while retention after 3 months measures long term engagement. The 3-month length was chosen with the average completion time for most nanodegrees being 6-12 months. It’s a long enough period that students are committed financially but also short enough as to not miss any students who learn at a much faster pace. The final total students with one payment metric is important because a rewards program would theoretically cost the company something financially and would not be sustainable unless enrollments are growing. The invariant metric would be number of user-ids enrolled. The user-ids will be randomized between the test and control groups as a result of also being the unit of diversion. Unlike our main study here, the follow-up study focuses specifically on students who have entered the free trial period of the nanodegree program and after. Every student has their own specific user-id, so each can only be counted once.   Finally, it’s worth noting this experiment would take significantly longer to collect data.