P7 - A/B Testing Final Project

Udacity is considering a change in their user interface. New users who attempt to enroll in a free trial will be prompted to answer a question about how much time they are willing to spend on their coursework. Users who enter a low number of hours available are encouraged by the system to access the course materials for free without enrolling.

This change is to “set clearer expectations for students up front, thus reducing the number of frustrated students who left the free trial because they didn’t have enough time”. The change should also avoid “significantly reducing the number of students to continue past the free trial and complete the course”.

Unfortunately we do not have enough time (or data!) to measure those who complete their courses. Instead, we will measure those who continue to use Udacity past their free trial period and make an initial payment for the service.

The null hypothesis is that enrollment (beginning the free trial) and retention (making the first payment) will not be affected by this change.

The alternative hypothesis is that the change will affect either the enrollment or the retention. Let’s run this experiment and evaluate the results.

Experiment Design

I chose two invariant metrics to monitor in this experiment:

Number of cookies: The number of users should stay consistent across both control and experiment groups.

Number of clicks: The proposed change affects a user’s experience *after* they click the “Start Free Trial” button. Therefore the number of clicks should not change between control and experiment groups.

I originally chose three evaluation metrics for the experiment.

Gross Conversion - How many users start their free trial and enroll with their credit card details? Some users will be prompted to continue accessing the materials for free and not enroll if they do not have a lot of time for the courses. Gross Conversion measures the users who proceed and enroll in the free trial.

Retention - This metric measures how many users who enroll in free trial actually continue to use the service and make at least one payment. One purpose of this change is to retain students past the free trial into the paid program.

Net Conversion - How many users who click on “start free trial” enroll in the program and make their first payment? Changes to this metric will affect revenues at Udacity. Lower net retention means less revenue. This is an important metric to watch.

Since some users will be prompted to reconsider signing up for a free trial, I expect a decrease in gross conversion (enrollments). A small decrease (or increase) in this metric will be welcomed.

Retention and Net Retention are important metrics. One goal of this change is to reduce "the number of frustrated students who left the free trial because they didn't have enough time". This means that ideally the retention rate should increase.

However, how does this affect NET conversion or the conversion from those who click "start a free trial" to paid customers? Will changes in enrollments affect this number? Hopefully this number will stay the same or even increase.

Measuring Standard Deviation

The analytic standard deviations for the evaluation metrics are as follows:

Gross Conversion:	0.0202
Retention:	0.0549
Net Conversion:	0.0156

Analytical estimates may be suspect here because the unit of analysis (cookie) is not the same as the unit of diversion (click). The actual standard error of these results will likely be higher than the results of the analytical standard deviation.

Sizing

To estimate the number of pageviews required for preparing control and experiment samples for the experiment, I used a sizing calculator available here:

http://www.evanmiller.org/ab-testing/sample-size.html

I used a Bonferroni correction since the evaluation metrics are dependent on each other. If Gross Conversion (enrollments) goes down, then perhaps Net Conversion (paying customers) will also go down. Therefore, for the calculator I used an alpha of 0.02 (0.05 / 3 rounded) instead of 0.05. From the # of required samples, I used the rates in the baseline data to predict the number of cookies required to generate this data. Here are the number of required cookies from my calculations:

Gross Conversion:	825,350
Retention:	6,062,182
Net Conversion:	875,325

Udacity's website gets 40,000 unique cookies (users) per day. Even if 100% of Udacity's traffic is used, it would take 152 days (almost 5-months) to get enough data to verify the results of this experiment on Retention (% of enrollments who transitioned to paid customers). This is not practical.

Therefore during the sizing process, I decided to remove Retention from the evaluation metrics and focus on Gross Conversion and Net Conversion instead. With 875,325 cookies required, with 100% of Udacity's website traffic, it will take 22 days to get enough data to run this experiment. This is much more reasonable and will give us the net conversion rate (# of trial users who convert to paying customers) which is a similar metric.

This experiment would be slightly risky for Udacity. Although the goal of the experiment is to improve users' satisfaction with their experience, warning busy users against enrolling in the program could negatively affect enrollment and revenue. If management was concerned, we could check the data after the first couple of days and reduce the experiment group (or roll back changes) if there is a large enough impact. Unfortunately reducing the experiment group will slow down data collection and extend the period of the experiment to over a month.

Analysis

When comparing control and experiment samples, it's important to do a sanity check and see that invariant metrics have not changed between the two samples.

I calculated the ratio of control sample data clicks and pageviews to the entire (control and experiment) sample.I compared this rate (0.5006 for pageviews, 0.5005 for clicks) against the mean (0.5) and pooled standard error (0.0012 and 0.0041 respectively). The rates fall within the standard error therefore changes are not statistically significant. This is expected.

During analysis of the evaluation metrics, I noticed different results for Gross Conversion and Net Conversion.

Gross Conversion

The Gross Conversion rate decreased 0.02 (2%). That means that there are 2% fewer users completing their free trial enrollment after clicking the "Start Free Trial" button. The standard error here is 0.0098 so the negative change is statistically significant.

Although this is a negative change, one reason for this change was to avoid enrolling potentially frustrated users meaning those who do not have time to spend on the coursework. How will this affect the rate of cookies that transition to paying customers?

Net Conversion

Changes to Net Conversion are smaller. The net conversion rate decreased by 0.0049. Since the standard error for this metric is 0.0077, and 0 (no change) is within the upper bounds, this change is **not** statistically significant. We cannot say with 95% Confidence that this change caused a decrease in net conversions.

Sign Tests

To verify the effects of the experiment, I also ran a sign test on the daily results using a sign calculator available at this URL: http://graphpad.com/quickcalcs/binomial1.cfm

For the Gross Conversion data, I noticed that the experiment rates were higher on just 4 of the 23 days. Given that the probability of data being higher or lower is 0.5 (flip-of-a-coin) for no changes, this result is unlikely. The two-tail P value here was 0.0026 which is statistically significant.

The Net Conversion Rate Sign Test result was consistent with the effects mentioned above. The experiment net conversion rate was higher than the control group on 10 of the 23 days. This is much more consistent with the rate of a coin toss. The two-tail P value was 0.6776 which is not statistically significant. We cannot say with a 95% CI that the decreased conversion rate was due to the experiment.

Summary

I used the Bonferroni correction in my experiment because the Gross Conversion rate (rate that a user clicks on 'start a free trial' and enrolls in a course) and the Net Conversion rate (rate that a user clicks on 'start a free trial', enrolls in a course, and makes a payment) are related. Decreasing the enrollments means fewer people who can eventually make a payment.

The effect size hypothesis tests and the sign tests are consistent. Differences in both the effect size and sign test ARE statistically significant. The Net Conversion rate changes are much smaller and not statistically significant. This is true for both the daily and overall rates.

It would be nice if we could compare Retention Rates (the rate of those enrolled who make their first payment) in the experiment and control groups using the Retention metric. Unfortunately the page views required for this were too large for Udacity's daily traffic. However, even without this metric, we have monitored Gross and Net Conversion rates and noticed a significant decrease in Gross Conversion rate (enrollment) with only insignificant decreases in Net Conversion rates.

The alternative hypothesis is correct. The changes did affect one of our two evaluation metrics. I would recommend additional experiments before changing the interface permanently. See Below.

Follow-Up Experiment

I would like to continue running this experiment for an additional month and add two additional metrics:

1) Run a survey at the end of the experiment to ask paying users about their general satisfaction (1-10) with the product. Comparing this metric for the control and experiment groups would be an interesting indicator of how effective the change removed ‘potentially frustrated’ users.
2) Keep collecting retention data for a 2nd month of paying customers. How many ‘frustrated’ users are cancelling their enrollments after the first payment?

I would keep the unit of diversion the same so that data could be compared against existing and new users (starting their trial periods and first payments). I believe both metrics would provide improved clarity that could verify (or introduce uncertainty to) the results.

For these two metrics, my hypothesis would be as follows:

Null Hypothesis: Changes to the UI do not affect user satisfaction or retention rates for the second month.
Alternative Hypothesis: Changes to the UI significantly affect both user satisfaction and retention rates for paying customers through the second month.