Template Format

This template can be used to organize your answers to the final project. Items that should be copied from your answers to the quizzes should be given in blue.

Experiment Design

Metric Choice

List which metrics you will use as invariant metrics and evaluation metrics here. (These should be the same metrics you chose in the "Choosing Invariant Metrics" and "Choosing Evaluation Metrics" quizzes.)

Invariant Metrics: number of cookies, number of clicks, click-through-probability.

Evaluation Metrics: Gross conversion, 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.

First, for the invariant metrics. The change that is being tested will not be noticed by users until they click the “Start Free Trial” button. So the number of cookies to view the course overview page should be no different, since the course overview page will be the same as before, and the way a user has navigated to the course overview page will not have changed either. For number of clicks, we expect it be invariant for the same reason, there is no change noticeable until after the click is made, and so whatever conditions lead to clicking are no different between the tests. Finally, click-through-probability will remain invariant as it is merely a ratio of these other two invariants. And thus its reason for being invariant extends from those.

Now for evaluation metrics. The evaluation metrics chosen were gross conversion and net conversion. The hypothesis of the experiment that we are attempting to confirm is that the addition of the proposed “time commitment” prompt without significantly reducing the number of students who continued past the free trial and completed the course. As noted, this prompt is shown after clicking “Start Free Trial”. Gross conversion covers the difference in enrollment rate for students who are shown the time commitment prompt and whether they still continue on to complete checkout. Net conversion is at the opposite end of the process, where we determine whether students remain past the 14-day free trial and end up making a payment after also clicking “Start Free Trial”.

To launch, we would ideally want a reduction in Gross conversion while net conversion would stay the same or increase. We would expect less people to complete checkout because they might realize they don’t have the time to commit. But at the same time we don’t want this change to significantly influence how many people actually go on to make payments.

I'll note that retention was originally chosen as an evaluation metric until it was later discovered that it would require far too many page views or far too long a duration to accurately test. What it measures is similar to net conversion, the primary difference being the use of user-ids rather than cookies which may have helped determine more uniqueness, but which also made it require a larger sample.

Also number of user-ids was not included as an invariant for we would expect that the additional prompt would reduce the number of people who enroll. However, it also doesn't help much as an evaluation metric. Gross and net conversion are rates while user-id is a raw count. A raw count will not accurately reveal any differences between the two groups if the control and experiment groups sizes are different, which they almost certainly will be by some amount. Gross and net conversion control this possible difference.

Measuring Standard Deviation

List the standard deviation of each of your evaluation metrics. (These should be the answers from the "Calculating standard deviation" quiz.)

Gross conversion: 0.0202
Net conversion: 0.0156

For each of your evaluation metrics, indicate whether you think the analytic estimate would be comparable to the 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.

To decide whether we would need to evaluate the empirical variability rather than analytical variability, we consider the unit of analysis and diversion of our evaluation metrics. Our evaluation metrics both have a unit of analysis of unique cookies that clicked the "Start Free Trial" button. This matches our unit of diversion which is cookies, and so we can expect the empirical variability to be similar to our analytic variability.

I would expect there can be some significant seasonal variation in these metrics. For instance, perhaps there are students who look around at Udacity courses who think they have time to sign up for one at the beginning of a semester and then end up too busy and have to drop before or after making a payment. Or perhaps they sign around a holiday without thinking about the additional expenditures that arise around that time. Or maybe they sign up as part of a New Year's Resolution and decide they don't want to pursue it further.

However, these seasonal variations should affect both the control and experimental group. The only concern might be that the additional time commitment prompt might have an added effect for people under one of these other seasonal conditions, like a student who considers the impact on their time. Thus it would seem important to schedule this at a relatively neutral time, avoiding holidays, summer time and other events as much as possible. There isn't likely to be much of a difference in the influence of these conditions on Gross conversion and Net conversion since they both measure what happens after the change. The only difference that might persist is a holiday which left someone with less money than they were expecting, which might influence net conversion more than gross conversion.

If time was available, doing an A/A test might be good just to get a good idea of the variation in that time period. Luckily we should have a fairly large sample which can help smooth this out and give us a good idea of what the variation is and how comparable to an analytic estimate it is.

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. (These should be the answers from the "Calculating Number of Pageviews" quiz.)

No Bonferroni correction used.
Pageviews: 685325

(If retention were included, we would have needed 4741213 page views to reliably measure a change.)

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. (These should be the answers from the "Choosing Duration and Exposure" quiz.)

Fraction of traffic exposed: 1
Length of experiment: 18 days

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

The risk for the people who participate in the experimental group is minimal. The only additional information they have to now provide Udacity with is their estimation of free time in a week. This does not seem to reveal anything too personal, and it's clear to the user that this question is being asked. This change will not harm or raise the potential for harm for any user.

It does however seem there is some potential for risk to Udacity since the change we're making might potentially affect how many people end up paying money. But given that the change is likely to be fairly small, and diverting all traffic to the experiment will still mean that %50 of the revenue stream is unaffected, we can go for the shortest possible time and divert all of the site’s traffic.

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. (These should be the answers from the "Sanity Checks" quiz.)

Number of cookies
Confidence Interval: .4988 to .5012
Observed: .5006
Passes

Number of clicks on “Start Free Trial”
Confidence Interval: .4959 to .5041
Observed: .5005
Passes

Click-through-probability on “Start Free Trial”
Confidence Interval: .0812 to .0830
Observed: .0822
Passes

For any sanity check that did not pass, explain your best guess as to what went wrong based on the day-by-day data. **Do not proceed to the rest of the analysis unless all sanity checks pass.**

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. (These should be the answers from the "Effect Size Tests" quiz.)

Gross conversion
Confidence Interval: -0.0291 to -0.0120
Statistically Significant
Practically Significant

Net conversion
Confidence Interval: -0.0116 to 0.0019

Not statistically significant
Not practically significant

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. (These should be the answers from the "Sign Tests" quiz.)

Gross conversion
P-value 0.0026
Statistically significant

Net conversion
P-value .6776
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.

The Bonferroni correction would lead us to require more certainty in the differences in our metrics before we conclude they represent actual changes, i.e. true positives. That is, the Bonferroni correction reduces the occurrence of false positives, while necessarily also increasing the occurrence of false negatives. However, the occurrence of false positives in our case is not too bad of an event. In our case we want to be “statistically certain” that both gross conversion decreases and that net conversion has stayed the same. Thus, while it’s possible we might introduce a false positive for one of these to occur, it’s less likely that we’ll end up with a false positive in both cases. Or, put another way, since we are looking for the verification of both outcomes simultaneously, if we increase our rate of false negatives by using Bonferroni correction, we can significantly reduce our chances of finding a real effect if one occurs, since if just one of our desired outcomes fails, we consider the experiment to have failed. (A middle of the road answer in this case would suggest further study.)

Recommendation

Make a recommendation and briefly describe your reasoning.

I recommend further study or to move on if time does not permit further study. Our gross conversion rate did appear to decrease and was both practically and statistically significant. At the same time, net conversion did not appear to significantly change. These outcomes were what we desired. That is, if these changes are true and not statistical anomalies, this change likely does not hurt Udacity’s income while also reducing the number of users who sign up and drop the course before paying. This likely has the desired effect of reducing frustration which is better for Udacity's long term presence as a generator of educational content.

The reason for additional study is suggested is that it's not clear that net conversion really didn't change. The confidence interval that we came up with was heavily weighed to the negative side -0.0116 to 0.0019 which suggests the possibility that net conversion did in fact practically change in the negative direction (net conversion's practical significance boundary was a .01 difference) and the lack of positive values in the interval suggest this was close to being a statistically significant change. Thus, before we can launch this change we do need to know more about its effect.

It may just be better to move on with a different attempted intervention as well. The likely decrease in gross conversion seems quite low for the potential decrease in net conversion as well. Thus, given that it may not be worthwhile to spend more resources on this intervention, we could move on to trying something new.

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.

To improve upon this idea of making sure that users have the time for the courses they are taking and that they are committing time, I would like to experiment with reminder emails that are sent if a paid user hasn't logged a lot of time in their account in a given week or time frame. The goal would be to increase user engagement and further improve a user's experience so that they keep learning and don't later feel like they've wasted money since they let the course get away from them but kept paying for it.

I would design this as as a cohort so that we follow people under similar conditions, and make the unit of diversion user-id so that we made sure we were following the same person. My hypothesis would be that the reminder e-mails improve user engagement and ultimately lead to more satisfied customers who feel like their experience with Udacity was a good one.

To explore this, I would be interested to try and get a sense of how customer satisfaction and engagement changes for people. We could look at how often they log in and are active on the website. However beyond this, if we really want to nail down whether this leads to less frustration and a more positive experience for our user we'll have to come up with other metrics beyond their states. We would want to give them surveys to measure their own self-reported satisfaction just to see if we can figure out any differences in self-reports and this is the only access we have to their own internal experience with the product.

We could also follow their rate of progression through a course, like how many lectures they get through and how many lessons. We may find that people who are making stable progression are less frustrated and happier with their experience. To detect how many people are cancelling their subscription early, we could look at the rate of cancellations over how many user-ids sign up. Another metric might measure how far a user-id gets on average before cancelling, an uptick in this and an uptick in completion rate itself might be telling.

A good invariant metric would be how many user-ids sign up for the class. Users won't see the e-mails until after signed up and so these shouldn't be affected. Another invariant to check would be to make sure that the distribution of users according to demographic criteria has remained the same in the control and experiment group. For example, we shouldn't expect enrollment to suddenly spike to include higher percentages of users from France.

This would need a decent length of time to get a reliable idea, probably at least 30 days, depending on how much traffic we wish to divert to the experiment. We can likely divert most, since the addition of reminder e-mails does not pose a risk to users. If they already provide e-mails as part of sign-up, then there is also no additional information they have to provide.

Resources

http://www.evanmiller.org/ab-testing/sample-size.html
http://graphpad.com/quickcalcs/binomial1.cfm