A/B testing – Reducing early cancellations for Undacity

Long Wan, July 7th, 2016

1. Overview

At the time of this experiment, Udacity courses currently have two options on the home page: "start free trial", and "access course materials". If the student clicks "start free trial", they will be asked to enter their credit card information, and then they will be enrolled in a free trial for the paid version of the course. After 14 days, they will automatically be charged unless they cancel first. If the student clicks "access course materials", they will be able to view the videos and take the quizzes for free, but they will not receive coaching support or a verified certificate, and they will not submit their final project for feedback.

In the experiment, Udacity tested a change where if the student clicked "start free trial", they were asked how much time they had available to devote to the course. If the student indicated 5 or more hours per week, they would be taken through the checkout process as usual. If they indicated fewer than 5 hours per week, a message would appear indicating that Udacity courses usually require a greater time commitment for successful completion, and suggesting that the student might like to access the course materials for free. At this point, the student would have the option to continue enrolling in the free trial, or access the course materials for free instead.

The hypothesis was that this might set clearer expectations for students upfront, thus reducing the number of frustrated students who left the free trial because they didn't have enough time—without significantly reducing the number of students to continue past the free trial and eventually complete the course. If this hypothesis held true, Udacity could improve the overall student experience and improve coaches' capacity to support students who are likely to complete the course.

The unit of diversion is a cookie, although if the student enrolls in the free trial, they are tracked by user-id from that point forward. The same user-id cannot enroll in the free trial twice. For users that do not enroll, their user-id is not tracked in the experiment, even if they were signed in when they visited the course overview page.

2. Metric Choice

There are several potential metrics.

- **Number of cookies:** That is, number of unique cookies to view the course overview page. (d_{min}=3000)
- **Number of user-ids:** That is, number of users who enroll in the free trial. (d_{min}=50)
- **Number of clicks:** That is, number of unique cookies to click the "Start free trial" button (which happens before the free trial screener is trigger). (d_{min}=240)

- **Click-through-probability:** That is, 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)
- **Gross conversion:** That is, 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)
- **Retention:** That is, 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. (d_{min}=0.01)
- **Net conversion:** That is, 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)

2.1 Invariant Metric

Number of cookies, number of clicks and click-through-probability are chosen as invariant metric. There are chosen in order to make sanity check later that ensure the same distribution of experiment group and control group.

The standard of choosing invariant metric is that they should not change across experiment and control group. Only those behaviors happen after the question has been asked might change.

As the unit of diversion is cookie, number of cookies is chosen rather than number of user-ids. Whether there will be more cookies or not certainly has nothing to do with the change stated in hypothesis.

People’s viewing the page and clicking the button occurred before question, so number of clicks and click-through-probability will not change.

2.2 Evaluation Metric

Gross conversion and net conversion are chose as the final evaluation metric.

Different from invariant metric, evaluation metric usually change after implementing the change in hypothesis. One could choose one or more objectives as evaluation metric, however, they should be meaningful enough to measure the result.

In the experiment, we would like to measure whether the change would reducing the number of frustrated students who left the free trial because of limit of time while others remain the same.

Gross conversion, retention and net conversion are all fraction-like indicators. Theoretically, if Udacity notified that the study need at least 5 hours per week, those who will be not able to spare 5 hours per week would stop proceeding. Gross conversion is expected to decrease since number of total clicks will be the same but number of checkout will go down. Retention is expected to increase because only people having more than 5 hours per week left and they are more likely to hold the study load and choose to continue. Net conversion is expected to not decrease since this change seems to not affect those who finally choose to pay. They are reasonable evaluation metrics.

However, measuring retention required a huge number of pageviews so that it might be too long to collect them. So this one should be deleted. It will be analyzed in following parts.

2.3 Others

Number of user-ids makes for neither an invariant metric nor an evaluation metric. On the one hand, it would be affected by the change which was made before enrollment. So it cannot be an invariant metric. On the other hand, as a count number, it is highly dependent on the number of cookies or page views. Thus it does not follow normal distribution and cannot be selected as evaluation metric.

3. Measuring Variability

Before conducting tests, we should know distribution of each evaluation metric. Knowing variability of baseline data is so important that we can judge whether the final changes of evaluation metric occur by chance or not.

Here is the baseline data.

Unique cookies to view page per day:	40000
Unique cookies to click "Start free trial" per day:	3200
Enrollments per day:	660
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

Based on central limit theory, we can consider distribution of probability of enrolling(gc) and probability of payment(nc) follow normal distribution. Count their standard deviation and calculate the 95% confidence interval. One should notice that average unique cookies we use are 5000, one eighth of above baseline data. So when calculating standard deviation using following expression, while fraction data remain the same, count data should be divided by 8 each.

$$\mathrm{SE} = \sqrt{\frac{p(1-p)}{N}}$$

$$SE_{gc} = \sqrt{\frac{0.20625(1-0.20625)}{3200/8}} = 0.0202$$

$$SE_{nc} = \sqrt{\frac{0.1093125(1-0.1093125)}{3200/8}} = 0.0156$$

These two metric use cookies as unit of analysis which matches unit of diversion and are expected to be accurate. However, as for probability of payment which I will finally delete, I might take time to collect an empirical estimate of the variability, because the unit of analysis is unique user-id, which is different from unit of diversion.

4. Sizing

4.1 Choosing number of samples given power

We should know how many samples of experiment and control group we need to get a reliable result. Given that α = 0.05, β = 0.2, knowing the baseline probability and minimum detectable effect, we could count out the number of clicks every day for one group, not using Bonferroni correction.

$$N_{gc} = 25835$$

$$N_{nc} = 27413$$

As for acquiring the number of pageviews needed, extra calculation is needed. First of all, click-through-rate is 0.08, so number above should be divided by the rate. Then, the previous number should time 2 for two groups.

$$n_{gc} = 25835 * \frac{2}{0.08} = 645875$$

$$n_{nc} = 27413 * \frac{2}{0.08} = 685325$$

Plus, I will show how many pageviews are needed with respect to retention:

$$N_{re} = 39115$$

$$n_{re} = 39115 * \frac{2}{0.20625 * 0.08} = 4741213$$

Since we should choose the largest one as our sample size, 4741213 seems to be the right one. But it will take us too long to collect this number of data. See following analysis.

4.1 Choosing duration VS exposure

As we have 40000 unique cookies to view page per day, if required sample size is 4741213, we have to spend 119 days to collect enough data. It is too long. So we kicked retention off from the evaluation metric list.

Then comes 685325, the larger one of the rest two. We have to spend at least 18 days to collect enough data, which is acceptable. So gross conversion and net conversion are the final evaluation metric here.

Spending too long on this experiment will be risky for a growing company. It is not acceptable economically. It is also not good for efficiency, as this market is quite competitive so that you should made decisions quicker to grasp the market. Taking too long cannot ensure that result would not be affected by other changeable factors throughout time, either. So maybe we will conclude that changing the button is effective, but actually it is affected by other things, like a new advertisement of Udacity on the way of testing and we will make wrong decisions. Therefore, I prefer diverting all Udacity's traffic to this experiment and finish it as soon as possible.

What's more, the experiment has its own risks. Some users might get mental hurt because they are "suggested" not to proceed if they don't have 5 hours per week.

This experiment is not trying to collect sensitive data, like financial status, marital status, political attitudes, so there are few risks on personal privacy.

5. Analysis

5.1 Sanity Checks

After gaining data of both experiment and control group, we should firstly measure whether invariant metric between two groups are truly invariant. Using 0.5 as the expected rate of diversion, calculate standard deviation of each metric and get the interval at 95% confidence level.

$$SE_{cookies} = \sqrt{\frac{0.5 * 0.5}{345543 + 344660}} = 0.0006$$

$$SE_{clicks} = \sqrt{\frac{0.5 * 0.5}{28325 + 28378}} = 0.0021$$

$$p_{ctr} = 28378/345543 = 0.0821$$

$$SE_{ctr} = \sqrt{\frac{0.0821 * (1 - 0.0821)}{345543}} = 0.000467$$

Considering critical value of 95% confidence level for normal distribution, we can calculate upper and lower bound by counting 1.96*SE as its marginal error and their means.

$$CI_{cookies} = [0.5 - 0.0012, 0.5 + 0.0012] = [0.4988, 0.5012]$$

$$CI_{clicks} = [0.5 - 0.0041, 0.5 + 0.0041] = [0.4959, 0.5041]$$

$$CI_{ctr} = [0.0821 - 0.0009, 0.0821 + 0.0009] = [0.0811, 0.0830]$$

Calculate observed value of each metric.

$$O_{cookies} = \frac{345543}{345543 + 344660} = 0.5006$$

$$O_{clicks} = \frac{28325}{28325 + 28378} = 0.5005$$

$$O_{ctr} = 28325/344660 = 0.0822$$

As we see, three metric lay in their own confidence interval, so we can regard experiment and control group as the same. Sanity check has been passed.

5.2 Check for Practical and Statistical Significance

We handle evaluation metrics for the next.

For gross conversion,

$$\hat{p} = \frac{3785 + 3423}{17293 + 17260} = 0.2086$$

$$SE = \sqrt{0.2086(1 - 0.2086)(\frac{1}{17293} + \frac{1}{17260})} = 0.004372$$

$$\hat{d} = \frac{3423}{17260} - \frac{3785}{17293} = -0.020555$$

$$ME = 1.96 * 0.004372 = 0.008569$$

$$CI = [-0.020555 - 0.00857, -0.020555 + 0.00857] = [-0.0291, -0.0120]$$

From the results above, as the upper bound, -0.012, is less than -0.01, gross conversion is both statistically and practically significant.

For net conversion,

$$\hat{p} = \frac{2033 + 1945}{17293 + 17260} = 0.1151$$

$$SE = \sqrt{0.1151(1 - 0.1151)(\frac{1}{17293} + \frac{1}{17260})} = 0.003434$$

$$\hat{d} = \frac{1945}{17260} - \frac{2033}{17293} = -0.004874$$

$$ME = 1.96 * 0.003434 = 0.006731$$

$$CI = [-0.004874 - 0.006731, -0.004874 + 0.006731] = [-0.0116, 0.0019]$$

From results above, both 0 and -0.0075 locate in the interval, so net conversion is neither statistically and nor practically significant.

5.3 Run Sign Tests

Then we go through sign test. Probability of success for one test is expected to be 0.5 and each test follows binomial distribution.

For gross conversion, 4 out of 23 samples have positive influences. Correspondingly, its two-tail p-value is 0.0026, which is smaller than 0.05. So it is statistically significant.

For net conversion, 10 out of 23 samples have positive influences. Correspondingly, its two-tail p-value is 0.6776, which is greater than 0.05. So it is not statistically significant.

If there were any discrepancy between sign test and previous results, we should take samples from more days and retest it. I think more days being considered, fewer chances to get discrepancy. If there are still discrepancy, maybe we did badly in sanity check and we should select control and experiment group, as well as better invariant metrics again.

5.4 Bonferroni Correction Interpretation

Bonferroni correction is not used throughout the analysis. This experiment requires that all expectations should be meet in order to pass the test. In this case, as the number of hypothesis increases, the likelihood of type two error goes up. However, Bonferroni correction is designed to handle type one error and decrease possibility of false positive, so it is unnecessary to use it in this project.

5.5 Make a Recommendation

Experiment results are not quite within expectation. By informing students of time requirement after clicking button, gross conversion goes down by 2% in average and it is both statistically and practically significant. Net conversion also decrease by 0.5% and it is not significant. However, the confidence interval does include negative figures, which means that it is possible that the new feature would have negative effects on the flow. That is not our expectation. So I will not recommend to launch it until further tests will be implemented.

The reason for possible decrease on payment may be that this change scare off those who have less than 5 hours per week but will not be frustrated by the study load, and those who have less than 5 hours per week but have the strong ability to study by nature.

6. Follow-up: how to reduce early-cancellations

The key point for reducing early-cancellations is that we should make students feel that Udacity is convenient to use and one can easily gain help when faced with problems. Rather than setting a time suggestion that may cause gross conversion to decrease even before they began to experience, we should encourage more experience and try to let the person feel safe during free trial.

Two resources that students can reach to solve problems are scheduling a coach and forum. One idea I come up with is to encourage students to ask at least 1 non-duplicated question on forum during free trial to earn a 20% discount for the first month. On one hand, the advantage of our quick and accurate and kind reply on forum can be showed to new students. On the other hand, it is helpful to enrich questions library on the forum. Students might be more willing to proceed to payment after fully experiencing the service. There are potential risks. Firstly, it will cause heavy load for forum administrators to integrate questions. Secondly, question quality is not guaranteed.

To ensure the question quality, this piece of message can be showed after enrollment and keep being showed on the bottom of screen during the period of free trial.

Hypothesis: it might encourage more student to experience the forum and get help, and finally increase the number of users who pay for the subscription.

Evaluation metrics: only retention is meaningful as the change only occurs during free trial period after enrollment.

Invariant metrics: Number of user-ids. They are not affected by the change.

Unit of diversion: user-id, because the prerequisite of the discount is having an account and having determination to start free trial. It makes the result more accurate than cookies do.