UDACITY P7: A/B Testing

Shawar Nawaz

Experiment Overview: Free Trial Screener

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.

METRIC CHOICE

Q. Which of the following metrics would you choose to measure for this experiment and why? For each metric you choose, indicate whether you would use it as an invariant metric or an evaluation metric.

Invariant Metrics – The criteria for choosing Invariant metrics is that the distribution should remain consistent across the control and experiment groups.

Choices:-

1) Cookies – The unit of diversion is a cookie. Thus should be a very good choice for an invariant metric. (d_{min}=3000)

2) Number of clicks – In this experiment, the user does not know that clicking the 'start free trial' button will lead to the screen pop up. Thus it should be a very good invariant. (d_{min}=240)
3) Click-through probability – For the same reasons as the number of clicks, this could also be a very good invariant. (d_{min}=0.01)

Evaluation Metrics – The metrics we want to evaluate. We should expect them to change across the control and experiment groups.

Choices:-

1) Gross conversion – Gross conversion in this case is the ratio of the number of people who decided to enroll in the free trial to the number of people who clicked the 'start free trial' button. I would expect people to see the warning pop-up, and not enroll if they feel like they wouldn't be able to sacrifice 5 hours a week to study. This number should decrease across the 2 groups and is a good metric to evaluate. (d_{min}= 0.01)
2) Retention – Of the people who decided to enroll for the free trial, only some of them may actually decide to pay. This could tell us if the pop-up had an effect on students who decided to continue past 14 days. But because of the pop up message, only more committed people may tend to join and reduce the number of users that enrolled, which in turn, increases the retention rate. Thus, this seems like a good metric to evaluate. (d_{min}=0.01)
3) Net conversion rate – People may click the start button, see the warning message and decide not to enroll. I expect this number to decrease over the experiment. Thus, this is a very good metric. (d_{min}= 0.0075)

Reasons for not choosing others:-

1) Number of user-ids – The number across the two groups are dependent on the users decision after seeing the warning message. Thus this may not be constant. Thus it doesn't make a good invariant metric. In terms of evaluation metrics, the others that I have chosen have a denominator and can be normalized. User ids on the other hand, cannot.

MEASURING VARIABILITY

For each metric you selected as an evaluation metric, estimate its standard deviation analytically. Do you expect the analytic estimates to be accurate? That is, for which metrics, if any, would you want to collect an empirical estimate of the variability if you had time?
Sample size = 5000 cookies

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

Std deviation (SD) = sqrt(p* (1-p)/n)

Thus SD(gross-conversion) = sqrt(0.20625 * (1 – 0.20625) * 40000 / (3200 * 5000))
= 0.02023

SD(retention) = sqrt(0.53 * (1 – 0.53) * 40000 / (660 * 5000))
= 0.0549

SD(net conversion) = sqrt(0.1093125 * (1 – 0.1093125) * 40000 / (3200 * 5000))
= 0.0156

Emperical standard deviation should be comparable to the analytical estimate for Gross conversion and Net conversion rate because they both have cookies as their unit of analysis, but cookies are also chosen as a unit of diversion.

For retention, we have number of user ids as the unit of analysis, which is not a unit of diversion, thus their empirical standard deviation will not be comparable to the analytical estimate.

SIZING

Using the analytic estimates of variance, how many pageviews **total** (across both groups) would you need to collect to adequately power the experiment? Use an alpha of 0.05 and a beta of 0.2. Make sure you have enough power for **each** metric.

Alpha = 0.05
Beta = 0.2

In cases like this, where the metrics we use are highly correlated to each other, it is normally not a good idea to use Bonferroni correction as it be very conservative towards it.

dmin(Gross conversion) = 0.01 ; base conversion rate (Gross Conversion) = 20.625%
dmin(retention) = 0.01; base conversion rate (retention) = 53%
dmin(net conversion0 = 0.0075; base conversion rate(net conversion) = 10.93125%

Therefore,

Sample size n (Gross conversion) = 25835

Sample size n (retention) = 39115
Sample size n (net conversion) = 27413

Therefore pages we'll need for each (both groups):-

Gross conversion -> 25835 * 40000 * 2/ 3200 = 645875

Retention -> 39115 * 40000 * 2/ 660 = 4741213

Net Conversion -> 25835 * 40000 * 2/ 3200 = 685325

Therefore, the max of these would be the number of pageviews required -> 4741213

CHOOSING DURATION AND EXPOSURE

What percentage of Udacity's traffic would you divert to this experiment (assuming there were no other experiments you wanted to run simultaneously)? Is the change risky enough that you wouldn't want to run on all traffic?
Given the percentage you chose, how long would the experiment take to run, using the analytic estimates of variance? If the answer is longer than a few weeks, then this is unreasonably long, and you should reconsider an earlier decision.

Average cookie views per day = 40000

Therefore the number of days it would take to get the number of pageviews we require = 4741213/40000 = 118.53.

This means that it would take about 119 days to reach the desired pageviews which is around 4 months. This ends up being too long a timeframe.

The second highest pageview requirement from the previous section is 685325. Here the number of days required = 685325/40000 =~ 18 days which is quite reasonable.

This would result in us dropping the retention metric. Looking back to the objective of the experiment, we see that we are trying to reduce the number of students that left the free trial because they didn't have time to work on the course.
Net Conversion is still a very good metric to measure that.

The experiment doesn't seem too risky to Udacity as we aren't really subjecting the students to any sort of harm, physical or mental. We are also not taking any private information from them or any sort of financial information either.

EXPERIMENT ANALYSIS

SANITY CHECKS

Start by checking whether your invariant metrics are equivalent between the two groups. If the invariant metric is a simple count that should be randomly split between the 2 groups, you can use a binomial test as demonstrated in Lesson 5. Otherwise, you will need to construct a confidence interval for a difference in proportions using a similar strategy as in Lesson 1, then check whether the difference between group values falls within that confidence level.

P = 0.5
Z factor for 95% CI = 1.96

Clicks :-

Total(control) = 28378
Total(experimental) = 28325
Total = 56703

SE = sqrt(0.5 * (1- 0.5) / (total)) = 0.0021

Margin of error = z * SE = 0.0021 * 1.96 = 0.00411
CI = 0.5 +/- margin of error = [0.4959, 0.5041]

Observed P = 28378/(28378 + 28325) = 0.5004

Sanity Check: PASS

Page Views:-

Total(control) = 345543
Total(experimental) = 344660
Total = 690203

SE = sqrt(0.5 * (1 – 0.5) / Total) = 0.0006

Margin of error = z * SE = 0.0006 * 1.96 = 0.00117
CI = 0.5 +/- margin of error = [0.4988, 0.5011]

Observed P = 345543/ (345543 + 344660) = 0.5006

Sanity Check: PASS

Practical and Statistical Significance

Next, for your evaluation metrics, calculate a confidence interval for the difference between the experiment and control groups, and check whether each metric is statistically and/or practically significance. A metric is statistically significant if the confidence interval does not include 0 (that is, you can be confident there was a change), and it is practically significant if the confidence

interval does not include the practical significance boundary (that is, you can be confident there is a change that matters to the business.)

Gross conversion:-

Enrollments(control) = 3785
Clicks (control) = 17293

Enrollments(exp) = 3423
Clicks (exp) = 17260

P = (3785 + 3423) / (17293 + 17260)

= 0.2086

SE = sqrt(p * (1- p) * (1/N1 + 1/N2)) ; where N is number of clicks

= sqrt(0.2086 * (1 – 0.2086) * (1/17293 + 1/17260))
= 0.004372

d = (3423/ 17260) - (3785/17293) = -0.0205

margin of error = SE * 1.96 = 0.004372 * 1.96 = 0.00856

CI = [d+/ margin of error]
= [-0.0291, -0.0119]

Since CI doesn't contain 0 or dmin value of 0.01, it is both practically and statistically significant

Net Conversion:-

Payments(Control) = 2033

Payments(Experiment) = 1945

p = (2033 + 1945) / (17293 + 17260)
= 0.115

SE = sqrt(0.115 * (1- 0.115) * (1/17293 + 1/17260))
= 0.00343

d = 1945/17260 – 2033/17293
= -0.00487

Margin of error = 1.96 * 0.00343
= 0.00671

CI = [-0.01158, 0.00184]

Since the CI contains 0 it is not statistically significant. And since the negative value of the dmin (-0.0075) in inside the CI, it is also not practically significant.

Bonferroni correction would not be ideal for this experiment as we are looking for gross conversion to significantly decrease as well as net conversion not to significantly decrease. If we were looking at just one of these, then we could use Bonferroni correction.

SIGN TEST

For each evaluation metric, do a sign test using the day-by-day breakdown. If the sign test does not agree with the confidence interval for the difference, see if you can figure out why.

Total Number of Days = 23

Gross Conversion :-

Positive change days = 4
2-tailed P value from calculator = 0.0026 < 0.025
Significant : Yes

Net Conversion:-

Positive change days = 10
2 tailed P value from calculator = 0.6776 > 0.025
Significant : No

Make a Recommendation

Finally, make a recommendation. Would you launch this experiment, not launch it, dig deeper, run a follow-up experiment, or is it a judgment call? If you would dig deeper, explain what area you would investigate. If you would run follow-up experiments, briefly describe that experiment. If it is a judgment call, explain what factors would be relevant to the decision.

I would recommend not launching the experiment.

The gross conversion reduced which is desirable and showed practical and statistical significant changes.

But Net conversion didn't show statistical or practical significant changes. Furthermore, when you look at the negative of the practical significance boundary, it falls within the CI. This may cause the risk of hurting udacity's business if this number went down by a considerable amount.

Follow Up Experiment

My follow up experiment would be pretty similar to the above experiment in the sense that I would try to focus on getting more students to continue beyond the two week trial. This could be done by creating some sort of incentives for the students to complete. During the two weeks, students could be constantly shown links to interviews of ex-udacity students who went on to make it big, or links to articles showing how the current course degree could lead to high paying jobs.

Null hypothesis: Showing links about how their degree will lead to high paying jobs to enrolled students will not create any significant difference to the net conversion.

Alternate hypothesis: Showing links about how their degree will lead to high paying jobs to enrolled students will increase the net conversion of students.

A good unit of diversion will be the user-id as each user will have his unique id once he enrolls.

A good invariant metric will be the number of user ids. This is because enrolled students reading links to high paying jobs will not deter a student from the 2 week trial and the number will remain constant across the two groups.

A good evaluation metric would be retention (ratio of users who made atleast one payment to students who decided to try out the program. An practically and statistically significant increase in this ratio would be a great sign to try out the new changes.