Final Project

March 7, 2016

1 EXPERIMENT DESIGN

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.

1.1 Metric Choice

- Number of cookies: That is, number of unique cookies to view the course overview page. (dmin=3000)
- Number of user-ids: That is, number of users who enroll in the free trial. (dmin=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). (dmin=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. (dmin=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. (dmin= 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. (dmin=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. (dmin= 0.0075)

1.1.1 Invariant Metrics

There are several invariant metrics that could be used over the course of this experiment. In order to be fit, an invariant metric should not change across experimental and control groups. After conducting the experiment, these will provide a way to double-check the integrity of our design. Because the screener pops-up after clicking on the 'start free trial' button, the number of pageviews, clicks, and the click-through-probability should remain unchanged during the experiment. Anything after the screener; number of user-id's, gross conversion, retention, and net conversion could be affected. Therefore, the invariant metrics chosen for this experiment are:

- Number of cookies (approximation of unique pageviews)

- Number of clicks
- Click-through-probability

1.1.2 Evaluation Metrics

Evaluation metrics are expected to change over the course of the experiment. By comparing differences between the control and experimental groups, we can measure the effect of the screener and test our hypothesis. Of the remaining metrics not considered to be invariant, user-id is excluded from the list of potential evaluation metrics. This is because user-id alone is a count, and gross conversion is a fraction that incorporates user-id while also offering a better way to track the effect of the screener. The evaluation metrics for this experiment are:

- Gross conversion rate (could measure whether or not the screener had an effect on enrollment)
- Retention rate (could measure whether or not the screener had an effect on the 14-day dropout rate)
- Net conversion rate (could measure whether or not the screener had any effect on the 14-day completion rate, although not able to tell us where in this process)

If the hypothesis is correct, we would expect to see specific changes in the evaluation metrics. Gross conversion would be lower as those students likely to drop during the 14-day trial would be filtered by the screener. Retention rate would be higher as those likely to drop would not have enrolled, and those who enrolled would not be likely to drop. Last, net conversion would be unchanged as the amount of students to continue past the free trial and eventually complete the course would have been unaffected.

1.2 Measuring Standard Deviation

Before conducting the experiment, data was collected to get daily values for cookies, enrollments, click through probability, gross conversion, retention, and net conversion on Udacity's website. The data collected is referred to as the baseline.

In the experiment, we predict that we will need approximately 5,000 cookies per day in each group. From this, a rough estimate of the expected standard deviation for each evaluation metric can be calculated. First, to get an approximation of the number of clicks and enrollments for this daily sample of 5,000 cookies, we scale by the fraction of pageviews in the sample over the pageviews in the baseline $\frac{5,000}{40,000} = 0.125$. Therefore, from 3,200 clicks and 660 enrollments in the baseline, we predict 400 clicks and 82.5 enrollments per day in the sample.

The number of clicks and enrollments follows a binomial distribution, and by the central limit theorem, the distrubution of the rates (gross conversion, retention, and net conversion) is gaussian. The standard deviation of these normally distributed rates is $\sigma = \sqrt{\frac{p(1-p)}{n}}$. The

rates for the evaluation metrics are:

$p_{gc} = 0.20625$
$p_r = 0.53$
$p_{nc} = 0.1093125$

and calculation of the standard deviations yields:

$\sigma_{gc} = \sqrt{\frac{0.20625(1-0.20625)}{400}} = 0.0202$
$\sigma_r = \sqrt{\frac{0.535(1-0.53)}{82.5}} = 0.0549$
$\sigma_{nc} = \sqrt{\frac{0.1093125(1-0.1093125)}{400}} = 0.0156$

Post experiment, the actual number of cookies used per day was higher than our estimated value. There were nearly 10,000 cookies per day per group rather than 5,000. Doubling the expected sample size for clicks to 800 and enrolls to 165, we can make a revised analytic estimate for standard deviation. The values that we expect to see are:

$\sigma_{gc} = 0.0143$
$\sigma_r = 0.0388$
$\sigma_{nc} = 0.0110$

Of the 3 evaluation metrics, only the analytically calculated standard deviation of retention is likely to match the empirical standard deviation seen in the experiment. This is due to the fact that the units of diversion and analysis are the same. Although gross conversion and net conversion also have user-id as the unit of analysis, they have cookies as the unit of diversion. This implies that our analytically calculated standard deviation could vary from the empirical standard deviation for these two metrics.

1.3 Sizing

1.3.1 Number of Samples vs. Power

To know the exact number of pageviews required for our experiment, we calculate the sample size that we will need for each evaluation metric. We allow a type I error rate of $\alpha = 0.05$, and type II error $\beta = 0.20$. The minimum detectable effect for each evaluation metric has been prespecified (as a business decision):

$d^{gc}_{min} = 0.01 \quad d^{r}_{min} = 0.0075 \quad d^{nc}_{min} = 0.01$

Using the rates from the baseline sample along with this α and β, a sample size calculator will give us the required number of samples for each evaluation metric. Each metric has it's own unit of size (clicks or enrolls), so once we arrive at the required sample size, we need to scale

from the given unit to pageviews by the ratio seen in the baseline. Finally we need to account for both groups in the experiment.

ratio of pageviews to clicks = 0.08
ratio of pageviews to enrolls = 0.0165

n_{gc} = 645,875 n_r = 4,741,213 n_{nc} = 685,275

The largest sample size is our limiting factor (retention rate), so we require a total of 4,741,212 pageviews to conduct the experiment.

1.3.2 Duration vs. Exposure

Considering the required pageviews, an exposure can be specified based upon the risk of the experiment, and from this a duration can be calculated. The exposure is dependent upon the risk involved and because the screener is a mild reminder about time commitment, it constitutes minimal risk. None of the participants could suffer physical harm as a result of the experiment, nor is sensitive data being collected, therefore a 100% exposure is a safe. Dividing total pageviews by the number of pageviews per day in the baseline (40,000), gives us a duration of 119 days were Udacity to divert it's entire traffic. This is too long of an experiment and we should reduce the duration. We can exclude retention as an evaluation metric and consider the next limiting metric, net conversion. With a revised 685,275 necessary pageviews, it would then take 18 days to run the experiment.

Excluding retention as a metric still allows us to test our hypothesis with net conversion. The two metrics are highly correlated and reiterating the objectives will make this relationship clear. The intention of the screener is to reduce the amount of people who enroll that would otherwise drop-out during the 14-day trial, while unaffecting the number of people to go on past the 14-day trial. Retention measures the difference in the rate at which people drop from enroll to completion of the 14-day trial using user-id as both the unit of diversion and analysis. Net conversion also uses the number of user-id's to complete the trial as the unit of analysis and therefore captures the effect of retaining students. With a minimum detectable effect of 0.0075, the constraint on net conversion allows us to ensure that the same amount of students still go on to complete the trial.

2 Experiment Analysis

2.1 Sanity Checks

Having conducted the experiment, we can double-check our invariant metrics to see if our underlying assumptions are being met. We expect that cookies and clicks are divided evenly between the control and experimental groups. Using an expected rate of diversion of 0.5, we

can calculate the standard deviation (using the a normal approximation to the rate as before), and construct a 95% confidence interval around our expected value. Comparing the observed rate, we can check if these two invariant metrics are reliable.

p = 0.5
α = 0.05
Z = 1.96

$\sigma_{cookies} = \sqrt{\frac{0.5(1-0.5)}{345543+344660}} = 0.00601841$
Margin of Error = 1.96 * 0.00601841 = 0.001179608
Confidence Interval = [0.4988, 0.5012]
Observed rate = 0.500639666

$\sigma_{clicks} = \sqrt{\frac{0.5(1-0.5)}{28378+28325}} = 0.002099747$
Margin of Error = 1.96 x 0.002099747 = 0.004115504
Confidence Interval = [0.4959, 0.5041]
Observed rate = 0.500467347

Both cookies and clicks pass the sanity check. For click-through-rate (CTR), we should observe more or less the same value across groups. Using the observed rate in the control group, we can construct a confidence interval, but instead compare the observed rate in the experimental group. This will test whether or not the two rates come from the same population which is what we would expect.

CTR_c = 28378 / 345543 = 0.082125813
$\sigma = \sqrt{\frac{0.0821(1-0.0821)}{345543}} = 0.000467$
Margin of Error = 1.96 * 0.000467 = 0.00091532
Confidence Interval = [0.0811, 0.0830]
CTR_x = 28325 / 344660 = 0.0822

CTR passes the sanity check, and all of our invariant metrics appear to be well chosen.

2.2 Results Analysis

2.2.1 Effect Size Tests

For each evaluation metric, we test for statistical and practical significance (whether or not the size of the effect is relevant from a business standpoint). The minimum detectable effect is the smallest difference that we will accept between experimental and control groups in order to be practically significant. Using the data collected, we calculate the rate in experimental and control groups for each evaluation metric (gross conversion, net conversion), and then define a new variable that is the difference between the rates (experiment - control). Using this newly defined variable, we construct a confidence interval which will then

set a range for the expected difference. Here we do not use the Bonferroni correction because we require that each of our evaluation metrics be significant. We are not basing our decision on the significance of one metric.

$\alpha = 0.05$
Z = 1.96

Gross Conversion

$r_c = 0.2188746892$
$r_x = 0.1983198146$
$\hat{d} = -0.020554874$

$Var_c = \frac{0.2188746892*(1-0.2188746892)}{17293} = 9.88657605 * 10^{-6}$
$Var_x = \frac{0.1983198146*(1-0.1983198146)}{17260} = 9.211417482 * 10^{-6}$
$Var_{\hat{d}} = 9.88657605 * 10^{-6} + 9.211417482 * 10^{-6} = 1.909799252 * 10^{-5}$

$\sigma_{\hat{d}} = 0.004370125$
ME = $8.5652 * 10^{-3}$
CI = [-0.0291, -0.0120]
d_{min} = (-)0.01

Net Conversion

$r_c = 0.1175620193$
$r_x = 0.1126882966$
$\hat{d} = -0.004873723$

$Var_c = \frac{0.1175620193*(1-0.1175620193)}{17293} = 5.999027983 * 10^{-6}$
$Var_x = \frac{0.1126882966*(1-0.1126882966)}{17260} = 5.793142782 * 10^{-6}$
$Var_{\hat{d}} = 5.999027983 * 10^{-6} + 5.793142782 * 10^{-6} = 1.179217076 * 10^{-5}$

$\sigma_{\hat{d}} = 3.43397029 * 10^{-3}$
ME = $6.7228 * 10^{-3}$
CI = [-0.0116, 0.0018]
d_{min} = (-)0.0075

Gross conversion is both statistically and practically significant. Net conversion is not statistically significant but the negative value of the minimum detectable effect is within the range of the confidence interval. To assess the practical significance, recall that our objective is to see no change in net conversion. We want the lower bound of the confidence interval to be

more negative than our minimum detectable effect which should be excluded from the confidence interval. Therefore net conversion is not practically significant.

2.2.2 Sign Tests

To further test each of our evaluation metrics, we can conduct a binomial sign test. Each day of the experiment is evaluated to see if there is a positive or negative difference across groups (experimental - control). We count each positive difference as a success, and each negative difference as a failure. Comparing the resulting p-values for each metric, we can determine significance. Gross conversion rate has 4 of 23 successes for a two-tailed p-value of 0.0026. This is much smaller than our individual type I error (from the Bonferroni correction) of 0.025 indicating statistical significance of gross conversion. Net conversion has 10 of 23 successes and a two-tailed p-value of 0.6776 indicating that net conversion is not statistically significant.

2.2.3 Summary

The effect size tests determine that gross conversion is both statistically and practically significant, while net conversion is neither. The gross conversion rate dropped in the experimental group by approximately 2% and thus the screener proved to be effective at reducing the number of students that enrolled from initial click. This supports our hypothesis if the screener is to be effective. Net conversion, however, was reduced by approximately 0.5% indicating that the screener had a negative effect on the number of students that would go on to complete the 14-day trial. In other words, the screener deterred some students from enrolling that would have otherwise completed the trial. This is not our intended effect and does not support the hypothesis.

The Bonferroni correction was not used in the analysis phase because our launch decision is based upon the significance of two metrics rather than just one. Had we used just one metric out of several to base our launch decision, the Bonferroni method would be appropriate. But, because the nature of our hypothesis requires that two effects be considered, we cannot base our decision in one metric alone.

The sign tests allow for an additional form of analysis. The conclusion from the sign test mirrors that of the effect size test, that gross conversion is significant but net conversion is not. Had we any discrepencies with regard to the signifance of the evaluation metrics between the sign and effect size tests, further study would be warranted. In this case, both tests agree and our conclusions with regard to both metrics are strongly supported.

2.3 Recommendation

The screener proved to be effective at reducing the number of people to continue from click to enroll, but it was not successful an unaffecting the number of students that would continue on past the 14-day trial. In fact, the screener appeared to increase the rate at which people

left the 14-day trial. Based upon this evidence, we should not launch the change to Udacity's website.

3 Follow-Up Experiment

A follow-up experiment could be based upon motivation with only a slight change from the previous experiment. It would require a method to approximate the number of hours that each student dedicated to the material in the first week. If a student committed less than the recommended number of hours, a message would pop-up upon login before the start of the second week to motivate the student to commit more time. Along with this message could be links to success stories or video of interviews with previous students who have completed that particular class or nanodegree. It could include information about where they went on to land a new job with their newfound skills from that class or nanodegree. Those that met the recommended number of hours wouldn't get a direct message but could still access this repository of interviews from their course homepage under the 'Resources' tab.

My hypothesis is that the message would motivate some students who might otherwise drop out during the 14-day trial to continue past and possibly complete the course. It would also not affect those people that would otherwise continue through the trial and complete the course had there been no pop-up message. In this case, the overall student experience in the forums could be more energized and improve beyond the first week, and coaching resources would be used on more enthusiastic and dedicated students.

Considering this design, retention rate would be the best way to test our hypothesis. Being that we would divert traffic evenly among the control and experimental groups, the most suitable invariant metric would be the number of user-id's to complete checkout and enroll in the course. It would be practical at this point, to divert students into the control or experiment group. It's worth noting that this experiment would likely take longer to conduct, recalling that the inclusion of retention rate added significantly more time to the duration than gross or net conversion.