UDACITY, UD257 A/B TESTING

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 the experiment. In order to be fit, they should not change between the experimental and control groups. After conducting the experiment, they 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 click-through-probability should remain the same across groups. Anything after the screener; number of user-id's, gross conversion, retention, and net conversion could be affected. The invariant metrics chosen for the experiment are:

- Number of cookies (approximation of unique pageviews)
- Number of clicks
- Click-through-probability

1.1.2 Evaluation Metrics

We expect that our evaluation metrics will change over the course of the experiment. By comparing differences between the control and experimental groups, we can measure the effect of the screener. 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 incorporates user-id while also offering a better way to track the effect of the screener during the experiment. The evaluation metrics to be used:

- 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 our hypothesis is correct, we would expect to see a higher retention rate for the experimental group. Conversely, gross conversion would be lower as those students likely to drop during the 14-day trial would be filtered by the screener. Net conversion may not change at all even if the screener is effective. To illustrate this, let's assume that 100 people clicked the 'start-free-trial' button in both the control and experimental group. In the control group, 10 people decided not to enroll, and 20 more dropped-out before the end of the trial period. In the experimental group, 25 people decided not to enroll after seeing the screener, and only 5 dropped-out before the end of the trial period. Although the screener was effective in filtering out potential drop-outs, net conversion across groups would be the same. Despite this possibiity, if the number of cookies is larger for the experimental group, it is possible that net conversion is higher. An increase would indicate that the screener had some effect to increase commitment past the 14-day trial. Because net conversion may not recognize an effect, it's reliability as an evaluation metric is questionable.

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 $\frac{5000}{40000} = 0.125$. Therefore, 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:

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

and calculation of the standard deviations yields:

$\sigma_{gc} = 0.0202 \quad \sigma_r = 0.0549 \quad \sigma_{nc} = 0.0156$

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 and enrolls, we can make a revised analytic estimate for standard deviation. The values that we expect to see are:

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

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 diversion, they have cookies as the unit of analysis. 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. Let us assume we allow a type I error rate of $\alpha = 0.05$, and type II error $\beta = 0.20$ for each metric. 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} = 645875 n_r = 4741213 n_{nc} = 685275

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

Given the required pageviews for our experiment, we can specify an exposure and calculate duration of the experiment. 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 an experiment. In order to cut down on duration, we can exclude retention rate as an evaluation metric and consider the next limiting metric, conversion rate. With 685,275 necessary pageviews, we can specify exposure at 50% of traffic. Diverting half the traffic per day (20,000), it would then take 35 days to run the experiment which is a reasonable duration. Because the screener is a mild reminder about study time, it constitutes minimal risk. There is no possibility that any of the participants suffer physical harm as a result of the experiment, nor is sensitive data being collected, therefore a 50% exposure is a safe.

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

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

$\sigma_{clicks} = \sqrt{\frac{p(1-p)}{28378+28325}}$ = 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. Because we are using multiple evaluation metrics to test our experimental hypothesis, we use the Bonferroni correction to determine the individual type I error for each evaluation metric:

$\alpha_{ind} = \frac{\alpha_{total}}{n}$ = 0.025
$\alpha_{ind/2}$ = 0.0125 (two-sided test)
Z = 2.24

Gross Conversion

r_c = 0.2188746892 r_x = 0.1983198146 $\widehat{d}$ = -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}
$Var_{\widehat{d}}$ = 9.88657605 * 10^{-6} + 9.211417482 * 10^{-6} = 1.909799252 * 10^{-5}

$\sigma_{\hat{d}}$ = 0.004370125
ME = 9.7888 * 10^{-3}
CI = [-0.0303, -0.0108]
d_{min} = (-)0.01

<u>Net Conversion</u>

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 = 7.692099585 * 10^{-3}
CI = [-0.0126, 0.0028]
d_{min} = (+)0.01

Gross conversion is both statistically and practically significant but net conversion is neither.

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 significant and net conversion is not. 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 enroll from click. This aligns with our hypothesis if the screener is in fact effective. Net conversion, however, was reduced by approximately 0.5% indicating that the screener did not increase the retention of students beyond the free trial period from click. This is the opposite effect that we would expect from our hypothesis. Even if the change was positive, it is too small to be practically significant for a launch.

It would have been helpful in the experiment to include retention rate during the effect size analysis. This metric could indicate whether or not those students that enrolled after seeing the screener were more likely to finish the 14-day trial. Including this metric caused the duration of the experiment to go beyond a reasonable limit and it could not be used.

The Bonferroni correction was used in the analysis phase in order to make a conservative estimate of the individual type I error rates. A conservative estimate is needed as we base our launch decision upon the practical significance of all metrics. In other words, all metrics need to be significant in order to launch the change, and because we are using multiple metrics, the possibility of a type I error increases. Thus by minimizing the individiual type I errors of each metric using the Bonferroni correction, we can arrive at an overall type I error that is acceptable.

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 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 effective at filtering out those students that would later drop during the 14-day trial. To the contrary, the screener appeared to increase the rate at which people left the 14-day trial. This appearance is not entirely surprising as net conversion is correlated with gross conversion. Most importantly, because of our exclusion of retention rate, we did not have an accurate way to measure whether or not people who enrolled after seeing the screener were more likely to continue past the 14-day trial. Based upon this evidence and reasoning, we should not launch the change to Udacity's website.

3 Follow-Up Experiment

A follow-up experiment could be based upon motivation and engagement. The experiment would require a method to approximate the number of hours that each student dedicated to the material per week. If a student committed less than the recommended number of hours in the first week, a message would pop-up upon login before the start of the second week to motivate the student to commit more time. Those that met the recommended number of hours wouldn't get a message.

My hypothesis is that the message would motivate some students to continue past the 14-day trial and possibly complete the course, but also cause some students to drop-out before that start of the second week. If the first is true, we may be able to motivate students to change their habits. If the second is true, those students that may not complete the program would then leave. In this case, the overall student experience in the forums could improve beyond the second week, and coaching resources could be used on more dedicated students.

Considering this design, we could include an additional evaluation metric, the '7-14 day' retention rate. Both the units of diversion and analysis would be user-id as in retention rate. The overall retention could allow us to evaluate if the pop-up message had any effect. The '7-14 day' retention rate would evaluate whether or not the pop-up message had an effect to motivate students to continue past the 14-day trial period. Suitable invariant metrics for this experiment would be number of cookies, clicks on start-free-trial, user-id, and click through probability. This experiment would likely take longer to conduct, recalling that the inclusion of retention rate required significantly more time than gross or net conversion.