A/B testing- Create your own A/B test and evaluate its outcome

Aarthy C Vallur

Experiment Design

Metric Choice:

The metrics available for the experiment can be classified as follows:

Number of cookies: The number of unique cookies is the unit of diversion and hance an invariant metric.

Number of user-ids: This may change between the control and experimental groups and hence is not an invariant metric. It will also not serve as a good evaluation metric because it cannot be normalized to the unit of diversion.

Number of clicks: The number of unique clicks, similar to cookies, is an invariant metric. The unique clicks also happen before our experiment starts, which happens when the cookies are randomly assigned. So, the number of clicks has little chance of varying between the control and experimental groups, making it a good invariant metric.

Click-through-probability: Since both units (number of unique cookies and clicks) that are used to arrive at the click-through-probability are invariant metrics, this is also an invariant metric. Also, click-through-probability will not change between the control and experiment groups.

Gross conversion: This is a metric that can measure the outcome of the experiment, since it measures the number of user- ids that pass the required hours per week clause to enroll in a free trial. It is also a normalized metric. This is a metric for evaluation of outcomes.

Retention: Retention is a metric that measures the outcome of the experiment as it measures the number of users that remained enrolled after the trial period over those who started the trial. This is also a good evaluation metric. But after considering that the size and duration required for the experiment are impractical if retention is used as an evaluation metric, it was eliminated before launching the experiment.

Net conversion: Net conversion, is a step further up Gross conversion, in that it measures those that went on to enroll past the trial data based on the number of unique cookies to start the free trial. This measures the success rate of the experiment and is a good evaluation metric.

The invariant metrics I choose are Number of cookies, Number of clicks and Click-through-probability. The evaluation metrics I choose are Gross conversion, Net conversion and Retention. It should be noted that, since the number of enrollees are filtered based on the time they can spend on the course per week, all three may be reduced in the experimental group compared to the control group. That said, though the number of enrollees in the free trial could decrease in the experiment, the number of payments should not decrease for the best experiment outcome. Ultimately, in order to launch the experiment, the Gross conversion must decrease in the

experimental group compared to the control group, while the Net conversion must not decrease significantly between the control and experimental groups.

Measuring Standard Deviation:

The standard deviations for my evaluation metrics, given a sample size of 5000 cookies were calculated as follows:

1. Gross Conversion = $\frac{\#\ of\ enrollees}{\#\ of\ unique\ cookies\ to\ click} = \frac{660}{3200} = 0.20625$

   When n = 5000, # of unique clicks = 400

   Standard Deviation = $\sqrt{\frac{p(1-p)}{N}} = \sqrt{\frac{0.2063(1-0.2063)}{400}} = 0.0202$

2. Retention = 0.53

   When # of cookies = 5000, number of enrollments = $\#\ of\ unique\ cookies\ to\ click \times Gross\ Coversion$

   $= 400 \times 0.2063 = 82.5$

   Standard Deviation = $\sqrt{\frac{p(1-p)}{N}} = \sqrt{\frac{0.53(1-0.53)}{82.5}} = 0.0549$

3. Net Conversion = $\frac{\#\ of\ enrollees\ past\ trial}{\#\ of\ unique\ cookies\ to\ click} = Gross\ conversion \times Retention = 0.1093$

   When n = 5000, # of unique clicks = 400

   Standard Deviation = $\sqrt{\frac{p(1-p)}{N}} = \sqrt{\frac{0.1093(1-0.1093)}{400}} = 0.0156$

Both Gross conversion and Net conversion are normalized using the unit of diversion. So it is highly likely that the analytical and empirical Standard deviations are similar in estimate/ The same cannot be said about Retention. Hence, I will trust the analytical Standard deviations of Gross and Net retention to be quite accurate. But for Retention, if time permits, collecting the empirical standard deviation would be useful.

Sizing:

Sample size using power:

For a positive outcome to this experiment, we want the Gross conversion rate to decrease while the net conversion rate remains relatively unchanged. In other words, we want only people that fulfill the time criteria to enroll in the free trial while retaining most that enrolled to remain enrolled after 14 days (make at least one payment). Since our outcome is dependent on 2 events happening together, we do not need the Bonferroni correction.

Sample size was calculated using the online calculator with the given
$\alpha = 0.05$ (5%) and $\beta = 0.2$ (20%)
Base conversion rates and d_{min} for the evaluation metrics were used as follows

Gross conversion = 20.625 and 0.01
Retention = 53 and 0.01

Net conversion = 10.93 and 0.0075

Sample size for each group from calculator
Gross conversion = 25835 clicks
Retention = 39115 enrollments
Net conversion = 27411 clicks

From these, calculating the number of pageviews needed for each group

Gross conversion = $\left(\frac{Sample\ size \times \#\ page\ views\ per\ day}{Pageviews\ to\ enroll\ per\ day}\right) \times 2 = \left(\frac{25835 \times 40000}{3200}\right) \times 2 = 645875$

Retention = $\left(\frac{Sample\ size \times \#\ page\ views\ per\ day}{enrollments\ per\ day}\right) \times 2 = \left(\frac{39115 \times 40000}{660}\right) \times 2 = 4741212$

Net conversion = $\left(\frac{Sample\ size \times \#\ page\ views\ per\ day}{Pageviews\ to\ enroll\ per\ day}\right) \times 2 = \left(\frac{27411 \times 40000}{3200}\right) \times 2 = 685275$

Duration Vs Exposure:

If 3 evaluation metrics are used (Gross conversion, Net conversion and Retention):

Duration = $\frac{Page\ views\ needed}{Page\ views\ per\ day} = \frac{4741212}{40000}$ = 119 days

If 2 evaluation metrics are used (Gross conversion and Net conversion):

Duration = $\frac{Page\ views\ needed}{Page\ views\ per\ day} = \frac{685275}{40000}$ = 18 days

Duration of 119 days is too long and impractical, though 18 days is also quite long, it is still possible and more practical than 119 days. So I will choose a duration of 18 days with Gross conversion and Net conversion as evaluation metrics.

Given the long duration of the experiment, I would like to use an exposure of 1. I would like to divert all of Udacity's traffic to the experiment. Personal information is not involved much and the students who sign up spend at least 5 hours and possibly not more on the website. So, to really achieve our goal, 100% traffic will need to be diverted to the experiment for a duration of 18 days. It does not create any additional issues for either the website or the user. So this is doable.
The user data involved in launching the experiment does include sensitive personal information such as student identities including name, date of birth and email addresses, credit card or bank account details and so on. But these are routinely collected and with secure web site design, do not pose significant higher risks to the users nor do they have to be carefully checked with consents and regulations. Therefore, the experiment does not pose additional risk to the users and can safely proceed without consents.

Analysis

Sanity checks:

Performed sanity checks on the 2 invariant metrics I plan to use, namely, number of unique cookies and number of unique clicks on the “Start trial button”. With a Z- score of 1.96, calculated the margin of error and upper and lower bounds of the 95% confidence interval for the sanity check between the control and experimental groups.

Number of cookies:
Control group - 345543
Experimental group - 344660
Total, N = 345543 + 344660 = 690203
Pdiff = 0.5
$\overline{P}$ = $\frac{345543}{690203}$ = 0.5006

Standard Deviation = $\sqrt{\frac{0.5006\,(1-0.5006)}{690203}}$ = 0.0006
Margin of error, m = 1.96 $\times$ 0.0006 = 0.0012
Lower limit and Upper limits of 95% confidence interval = (0.5 - 0.0012, 0.5 + 0.0012)
= (0.4989, 0.5012)
$\overline{P}$, 0.5006 $\in$ (0.4989, 0.5012)

Number of clicks:
Control group - 28378
Experimental group - 28325
Total, N = 28378 + 28325 = 56703
Pdiff = 0.5
$\overline{P}$ = $\frac{28378}{56703}$ = 0.5004

SE = $\sqrt{\frac{0.5004\,(1-0.5004)}{56703}}$ = 0.0021
Margin of error, m = 1.96 $\times$ 0.0021 = 0.0041
Lower limit and Upper limits of 95% confidence interval = (0.5 - 0.0041, 0.5 + 0.0041)
= (0.4959, 0.5041)
$\overline{P}$, 0.5004 $\in$ (0.4959, 0.5041)

Click-thru probability:
Control group, clicks - 28378
Experimental group, clicks - 28325
Total clicks, N = 28378 + 28325 = 56703
Control group, cookies- 345543
Experimental group, cookies - 344660
Total cookies, N = 345543 + 344660 = 690203

Pdiff = 0.08212

$\overline{P}$ = $\frac{56703}{690203}$ = 0.08215

SE = $\sqrt{\frac{0.0821(1-0.0821)}{344660}}$ = 0.000468

Margin of error, m = 1.96 $\times$ 0.000468= 0.000917

Lower limit and Upper limits of 95% confidence interval = (0.08212- 0.000917, 0.08212+ 0.000917)
= (0.0812, 0.0830)

$\overline{P}$, 0.08215 $\in$ (0.0812, 0.0830)

Since the observed value falls within the the 95% confidence interval, all invariant metrics pass the sanity test.

Analysis of Results:

To conclude whether the results are statistically significant and practically significant to the business, calculated the upper and lower limits of the 95% confidence interval for the 2 evaluation metrics chosen- gross conversion and Net conversion through the duration of the experiment.

Bonferroni Correction : The Bonferroni correction was not used, since the outcome of the experiment is based on an additive effect- Decrease in Gross conversion **and** no decrease in Net conversion. We want both evaluation metrics to behave as expected. This is not a scenario where the Bonferroni Correction will help avoid errors. It is better used when in an either or scenario where Type 1 errors are expected. Hence the Bonferroni correction was not used in this experiment.

Gross Conversion:

Enrollment, Control - 3785

Enrollment, Experimental - 3423

# of clicks, Control - 17293

# of clicks, Experimental - 17260

$\overline{P}$ = $\frac{3785+3423}{17293+17260}$ = 0.2086

$\overline{d}$ = $\frac{3423}{17260}$ - $\frac{3785}{17293}$ = -0.0206

SE = $\sqrt{\frac{0.2086\ (1-0.2086)}{17293+17260}}$ = 0.0044

Margin of error, m = 1.96 $\times$ 0.0044 = 0.0086

Lower limit and Upper limits of 95% confidence interval = (-0.0206 - 0.0086, -0.0206 + 0.0086)
= (-0.0291, -0.01198)

Net Conversion:

Payments, Control - 2033

Payments, Experimental - 1945

# of clicks, Control - 17293

# of clicks, Experimental - 17260

$\overline{P} = \frac{2033 + 1945}{17293 + 17260} = 0.1151$

$\overline{d} = \frac{2033}{17260} - \frac{1945}{17293} = -0.0049$

SE = $\sqrt{\frac{0.1151\ (1-0.1151)}{17293 + 17260}} = 0.00343$

Margin of error, m = $1.96 \times 0.00343 = 0.0067$

Lower limit and Upper limits of 95% confidence interval = (-0.0049 - 0.0067, -0.0049 + 0.0067)
= (-0.0116, 0.0019)

Gross conversion is both statistically and practically significant, while Net conversion is not. Hence the experimental and control groups are significantly different when measuring Gross conversion but not so when measuring Net conversion.

Sign Tests:

I used this calculator as discussed in class to get 2-tailed P values for Gross and Net conversion

1. **Gross Conversion:**
   Total # of trials = 23
   Total successes= 4
   Probability = 0.5
   2-Tailed P-value from calculator = 0.0026
2. **Net Conversion:**
   Total # of trials = 23
   Total successes= 10
   Probability = 0.5
   2-Tailed P-value from calculator = 0.6776

In the case of Gross conversion, 0.0026 < 0.05 and the number of successes is also less than the probable successes (4 < 11.5). This is not the case with Net conversion, 0.6776> 0.005 and 10 is much close to 11.5 than 4.
So, the P- value suggests that Gross conversion is significantly different between control and experimental groups, while Net conversion is not.

Recommendation:

The outcome of this experiment must be to reduce the number of free trials that do not result in enrolment (payment) but not decrease the number of free trials that do end up in enrolment (payment). The productive outcome must reduce trials that are not fruitful, and in this way streamline coaching resources for only those users who enrol and commit to the course. The filter used is a time commitment requirement. We see that the Gross conversion, which measures the number of users that take up the free trial after satisfying the time commitment filter, does decrease

over the 18 day experiment. This is desirable and makes business sense. But we cannot clearly measure the effect on Net conversion, which measures the productive enrolment rate, which is the other goal of the experiment. Net conversion shows no significant change between experimental and control groups. The confidence interval of the Net conversion includes the lower limit of the practical significance boundary (-0.0075) but not the upper limit (0.0075). So there is a possibility that Net conversion actually decreases during the experiment. This would not be good for the business and goes counter to our initial hypothesis which is that, net conversion will not decrease as a result of the experiment. Unless we ascertain that net conversion does not decrease, launching this long experiment will be pointless. More data must be collected and possibly, other filters must be added to understand the effect of this experiment on Net conversion before it is launched.

Follow-Up experiment to reduce early cancellation:

In order to reduce early cancellation, I would suggest a follow- up experiment focussed on incentivising the student in enrollment. I have 2 strategies to help retain enrollees:

1. Adding an incentive to remain enrolled at the end of the trial period, such as targeted and personalized coaching. This could be in the form of asking the student if they want more help in an area they find difficult and following up, weekly one-on-one sessions with their progress. The promise of extra support personalized to their needs will more likely keep students enrolled.
2. Offering a marked discount if a student remains enrolled past a certain duration or pledges to complete the course in a certain duration.

Let us consider a follow- up experiment with the first incentive- Targeted and personalized coaching. Our experiment will consist of offering those that click on the “Start free trial” button, the incentive of free, personalized coaching complete with 15 minute weekly one- on- one sessions to discuss and help the student with that week’s lessons and quizzes, once they cross the 14- day free trial period. The control group will not get this incentive. This experiment will be performed after a user enrolls in the free trial.

Metrics:

The unit of diversion chosen is User-ids, which are a more stable metric than cookies. The only Invariant metric that can be used is the number of unique user-ids who are enrolled in the free trial, given the experimental design. This is also the unit of diversion and hence will not change between control and experimental groups. I want an evaluation metric that measure the early cancellation rate (Number of users that cancel after free-trial period over total number of users in free trial). The experiment is successful if cancellation rate decreases indicating drop in early cancellation. This would mean better retention, thereby keeping students that are motivated to pursue the course.