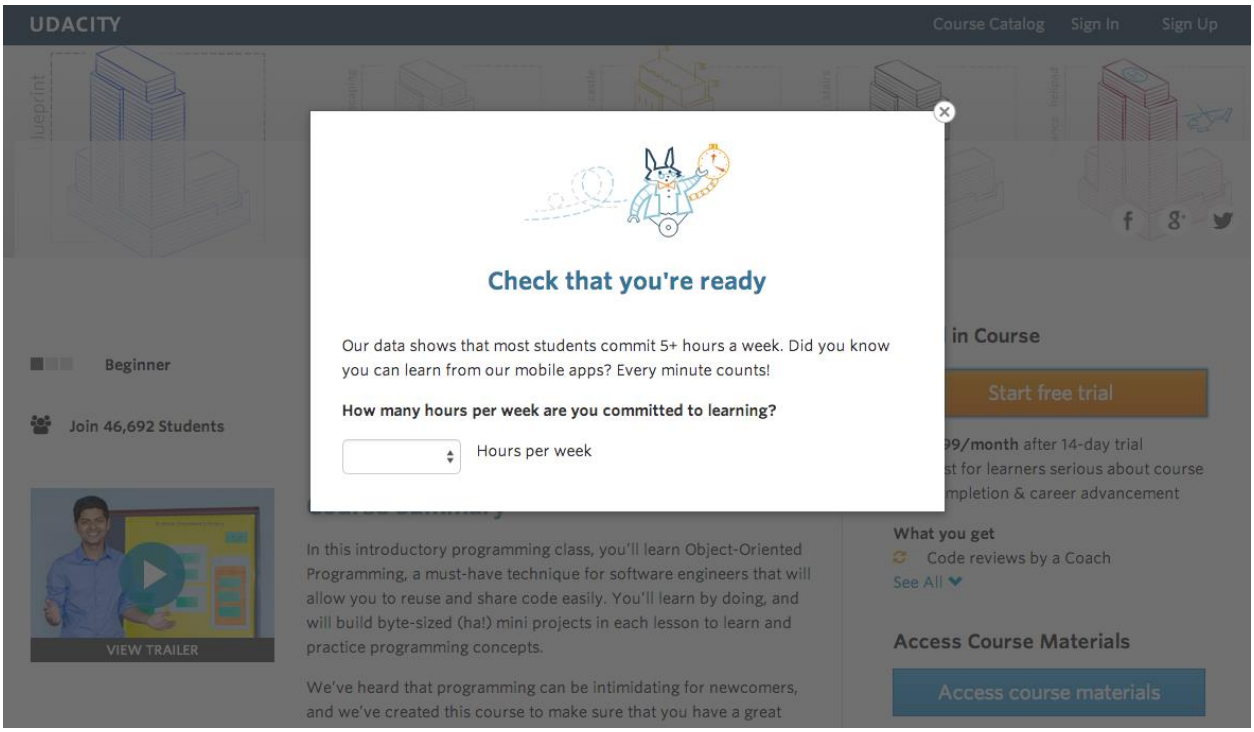
Data Analyst Nanodegree – Design an A/B Test

Introduction

In this experiment I will be A/B testing a change on Udacity's website, where if a student clicked "start free trial", they will be asked how much time they will be able to devote to the course and depending on whether they chose between greater than five hours or not, they will be directed to different workflows on the website. The student can make their choice through the screen below where they indicate the number of hours they can commit to learning.



The aim of this experiment is to test whether giving this kind of choice to the student can improve their experience if the student enrolled in a free trial period, which lasts for 14 days, and could not commit the required hours and are automatically charged at the end of the trial period. Being unaware of the demands of the course unless explicitly committed though this screen and then being charged might be leading to frustrated students. Udacity wants to test whether introducing this change can reduce the number of such frustrated students, raise their level of awareness regarding demands of the courses leading to greater success in course completion apart from an improved experience.

Experiment Design

Metric Choice

Chosen Invariant Metrics –
"Number of cookies", "Number of clicks" and "Click through probability"
Chosen Evaluation Metrics –
"Gross conversion", "Retention" and "Net conversion"

Invariant metrics can't play the role of evaluation metrics and vice versa. Below are the reasons for choosing these versus not choosing others:

- Number of cookies – since this is the number of unique cookies that visit the course overview page, the difference in site behavior between experiment and control groups occurs only after a student clicks the "start free trial" button. The number of unique cookies visiting the course overview page remains the same irrespective of the fact whether they later click the free trial button or not.
- Number of user-ids – This represents the number of users who enroll in the free trial. Since the enrollment to free trial happens only after clicking the "start free trial" button, hence this number would be affected by the experiment and cause variance between experimental and control groups. The reason for not choosing this variable as an evaluation metric is that the variable is not normalized and would lead to different distributions, further this feature is also captured in the gross/net conversion metric, which are also normalized variables.
- Number of clicks – This represents the number of users that click a trial button. Since a click happens just after visiting the course overview page and before display of the experimental screen hence it can be treated as an invariant metric.
- Click through probability – This metric represents is the ratio of number of unique cookies that click the "start free trial" button and number of cookies visiting the course overview page. Since the 2 variables are invariant, hence their ratio is also bound to be invariant.
- Gross conversion – This metric is the number of user-ids that complete the checkout and enroll in free trial divided by the number of unique cookies to click the "start free trial" button. This metric is directly measuring the students going beyond the new screen and enrolling for experimental group, which can be compared with the numbers for control group. Hence it's a good evaluation metric.
- Retention – This is the number of user-ids that remain enrolled after the 14 day free trial period divided by the number of user-ids that complete the checkout. This can also be seen as a ratio of Net conversion and Gross conversion. Since with the experiment both the numerator and denominator are expected to change, it can be difficult to understand the individual effect through this variable only. This is a kind of simple composite variable and hence I have made it a part of this experiment to understand how it moves. Later I discover that based on the baseline numbers provided the number of samples required for the experiment is prohibitively high for the experiment to finish in the stipulated time. Hence I removed this metric after Sizing and Duration sections.
- Net conversion – This metric is the number of user-ids that remain enrolled past the 14 day free trial period (making at least 1 payment) divided by the number of unique cookies to click the "start free trial" button. This metric is measuring whether there was a role of the new screen in changing the proportion of people going beyond the trial period when they were exposed to the new screen versus those who were part of the control group. Hence this serves as an evaluation metric.

In the evaluation metrics I will look for a decrease in Gross conversion in case of experimental group since now the student is made aware of the commitment and hence only the more serious students would sign up for the course. On the other hand the Net conversion, when comparing between experimental and control group should remain relatively similar, further there should be no decrease in net conversion as we don't want the revenues to be impacted negatively due to this change.

Measuring Standard Deviation

Based on the information provided in "Final Project Baseline Values" spreadsheet

- Number of cookies = 5000
- Number of clicks with click through probability of "start free trial" = 5000 × 0.08 = 400
- Number of enrollments, given clicks on “start free trial” = 5000 × 0.08 × 0.20625 = 82.5
- Standard Deviations
    - Gross Conversion = sqrt((0.20625 * (1-0.20625))/400) = 0.0202
    - Retention = sqrt((0.53 * (1-0.53))/82.5) = 0.0549
    - Net Conversion = sqrt((0.1093 * (1-0.1093))/400) = 0.0156

For both gross conversion and net conversion the unit of analysis (denominator) is number of unique cookies that click the “start free trial” button. The unit of diversion is also same for both and so the analytical (or theoretical) variability would be comparable to empirical variability. Hence both gross conversion and net conversion are expected to have similar distribution and standard deviation, but same can’t be said for retention. The unit of diversion is not used in case of Retention and hence empirical variability won’t be similar to analytical variability, we will have to estimate empirical variability separately.

Sizing

I did not use Bonferroni correction during the analysis phase because when multiple hypothesis are tested, the chance of a rare event and thus the chance of incorrectly rejecting a null hypothesis (making type 1 error) increases. But Bonferroni correction compensates for this increase in chance of making type 1 error by reducing the significance level to alpha/m (where m is the number of comparisons). Reducing alpha lowers the false positive rate but increases the chance of not rejecting a null hypothesis when alternative hypothesis is true (or making a type II error). Since in this experiment we need that all metrics must satisfy their respective criteria to launch the experiment and an increase in false negative (or type II error) will cause the experiment to not being launched even though it was valid or we might miss out on true positives. We are expected to reject the null hypothesis, but using Bonferroni might lead us to failing to reject the null hypothesis even if it is false (type II error chances being increased); further Bonferroni correction is not advised to be applied to variables that are inter-dependent. For the reasons stated above Bonferroni correction is not advised.

Number of Samples vs. Power

I used the online calculator from evanmiller.org to calculate the page views, the alpha is 0.05 and beta is 0.2 as explained in project instructions.

Gross Conversion – Using the baseline conversion rate of 20.625% and minimum detectable effect of 1% I get a sample size of 25,835. Multiplying with baseline unique cookies from page views (40,000), dividing by unique cookies clicking the “start free trial” (3200) and multiplying by 2 for both experiment and control groups, we get:
Required sample size = 25835 x 40000 / 3200 x 2 = **645625**

Retention – Using the baseline conversion rate of 53% and minimum detectable effect of 1% I get a sample size of 39,155. In this case enrollments per day is 660, while page views’ unique cookies are 40,000.
Required sample size = 39155 x 40000 / 660 x 2 = **4741212**

Net Conversion – Using the baseline conversion rate of 10.931% and minimum detectable effect of 0.75% I get a sample size of 27,413
Required sample size = 27413 x 40000 / 3200 x 2 = **685325**

For now we can go with the maximum sample size of 4,741,212 page views.

Duration vs. Exposure

For initial analysis I chose to consider 2 values, 60% and 100% of website traffic used for this experiment. The following table summarizes the days needed:

	60%	100%
sample = 4.7 million	198 days	119 days
sample = 685,325	29 days	18 days

Given that experiment has to be conducted within a few weeks, I chose the sample size to be **685,325** and traffic exposure to be 60% to conduct the experiment in **29** days or within a month, since a shorter duration might expose the experiment to some days that record abnormal activity, a longer duration would help even out the effects.
The experiment doesn't seem to be risky and not dealing with any sensitive information. It might give an enhanced experience and a pleasant surprise to someone who didn't view the new screen earlier in the experimental group, although might cost revenue loss at the expense of uninformed decision by the students to enroll in a course.

Experiment Analysis

Sanity Checks

Below are the 95% confidence intervals for the values I expect to observe, versus the observed value:

Number of Cookies –
Observed = 345543 / (345543 + 344660) = 0.5006
Std Err = sqrt (0.5 * 0.5 * (1/344660 + 1/345543)) = 0.0006
Margin of Err = 0.0006 * 1.96 = 0.0012
Lower = 0.5 – 0.0012 = 0.4988 | Upper = 0.5 + 0.0012 = 0.5012

Metric passed the sanity check

Number of Clicks –
Observed = 28378 / (28378 + 28325) = 0.5005
Std Err = sqrt (0.5 * 0.5 * (1/28325 + 1/28378)) = 0.0021
Margin of Err = 0.0021 * 1.96 = 0.0041
Lower = 0.5 – 0.0041 = 0.4959 | Upper = 0.5 + 0.0041 = 0.5041

Metric passed the sanity check

Click through Probability –
Observed = CTRexp – CTRcnt = 0.08218 - 0.082126 = 0.00006
CTRpool = [Clicks Cont + Clicks Exp] / [Pageviews Cont + Pageviews Exp]
= 56703 / 690203
= 0.082154
Std Err (pool) = sqrt (CTR_pool * (1 - CTR_pool) * (1/Pageviews Cont + 1/Pageviews Exp))
= sqrt (0.082154 * (1-0.082154) * (1/344660 + 1/345543))
= 0.00066
Margin of Err = 0.00066 * 1.96 = 0.00129
Lower = 0 – 0.00129 = -0.00129 | Upper = 0 + 0.00129 = 0.00129

Metric passed the sanity check

Result Analysis

Effect Size Tests

Confidence intervals for evaluation metrics at 95% confidence interval

Gross Conversion –

Prob = total enrollments / total clicks
Pcont = 3785 / 17293
= 0.21887
Pexp = 3423/ 17260
= 0.19831
Gross Prob pool = (3785 + 3423) / (17293 + 17260) = 0.2086
Std Err pool = sqrt (0.2086 * (1-0.2086) * (1/17293 + 1/17260) = 0.00437
Margin Err pool = 1.96 * 0.00437 = 0.00856
Difference = 0.19831 - 0.21887 = - 0.02056

Lower bound = -0.0291 Upper bound = -0.0120

Dmin for Gross Conversion as per instructions was 0.01, the metric seems to be statistically (didn't cross the 0) and practically significant (as it shows a 1.2% to 2.9% decline).

Net Conversion –

Prob = total payments / total clicks
Pcont = 2033 / 17293
= 0.11756
Pexp = 1945 / 17260
= 0.11268
Net Prob pool = (2033 + 1945) / (17293 + 17260) = 0.1151
Std Err pool = sqrt (0.1151 * (1-0.1151) * (1/17293 + 1/17260) = 0.00343
Margin Err pool = 1.96 * 0.00343 = 0.00672
Difference = 0.11268 - 0.11756 = - 0.00488

Lower bound = - 0.0116 Upper bound = 0.00184

Dmin for Net Conversion is 0.0075, the metric is not significant either practically or statistically (crossed 0)

Sign Tests

Done using http://graphpad.com/quickcalcs/binomial1.cfm

Gross Conversion –

Improvement days = 4
Total days = 23
Hypothesis probability = 0.5

p-value = 0.0026 (statistically significant as it is smaller than alpha of 0.05)

Net Conversion –

Improvement days = 10
Total days = 23
Hypothesis probability = 0.5
p-value = 0.6776 (not statistically significant)

Summary

I did not use the Bonferroni correction as already explained in the Sizing section above. To reiterate, since Bonferroni correction reduces the chance of Type I error and increases the chance of Type II error and so it might lead us to failing to reject the null hypothesis even if it is false; further Bonferroni correction is not advised to be applied to variables that are inter-dependent which is the case here.

Both the effect size test and sign tests led us to consistent results where Gross conversion is seen to be statistically and practically significant.

Recommendation

From the experiment we see a decline in gross conversion rate as expected but we also expected that there shouldn't be an effect on net conversion, which we can't say for sure as the confidence interval seems to go beyond the negative practical significance boundary of -0.0075. Hence launching the experiment might cause a revenue loss to Udacity and therefore I don't recommend launching the experiment as per the present design and should be further evaluated.

Follow-Up Experiment

As a follow-up experiment I would recommend a couple of mechanisms of informing students about their progress during the free trial course. If the student is spending the right amount of time taking the trial class they are expected to reach a certain milestone which may be passing a quiz or completing a lesson in the course. A solution could be engineered where in at the middle of the trial period, if the student has crossed the milestone they are encouraged about their progress. On the other hand if they are lagging behind, they must be shown a more attention grabbing screen reminding them of their lack of progress and giving them an option to cancel if they can't keep up or continue the course with an acknowledgement. If the student is not logging to the website frequently they could be reminded through a mobile app notification and an email with content being the similar as the pop-up screen. Keeping the student informed at an intermediate stage of the trial helps them set the expectations right and improve their experience with the Udacity platform.

The metrics for this experiment would be the users who acknowledge their continuing the program. We can also measure the students who cancel and let them submit reason for cancellation which is another source of a timely feedback before a frustrated student ends his class abruptly. Another metric would be retention i.e. the number of users who remain enrolled and make at least the first payment. The unit of diversion would be user-id of enrolled students as this experiment is for enrolled students and the invariant metric would also be the enrolled user-ids.