A/B Testing Final Project

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.

Metric Choice

Invariant Metrics:

An invariant metric should not change across the experimental and control groups. Therefore, the invariant metrics can be used to sanity check the integrity of the experimental design. In this experiment, because of the screener pops-up after clicking

the “Start free trial” button, the number of page views, Clicks and the Click-through-probability should not be affected by the experimental design. Therefore, they all can be considered as invariant metrics. Given that the Click-through-probability is actually calculated from the Number of clicks and the Number of cookies, I only chose the following two as invariant metrics:

- **Number of cookies**: number of unique cookies to view the course overview page.

- **Number of clicks**: number of unique cookies to click the “Start free trial” button.

These two invariant metrics should stay the same in order to launch the experiment.

Evaluation Metrics:

Evaluation metrics are expected to change over the experiment. Differences are expected to be observed between the experimental and control groups. After the screener pops-up, anything could be affected by the experimental design, including the Number of user-ids, Gross conversion, Retention and Net conversion. Given that Gross conversion, Retention and Net conversion are incorporated the number of user-ids, I chose the following three as evaluation metrics:

- **Gross conversion**: 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.

- **Retention**: 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.

- **Net conversion**: number of user-ids to remain enrolled past the 14-day boundary (and thus make at least one payment) divided by number of unique cookies to click the “Start free trial” button.

If the hypothesis is true, we would expect a decrease of Number of user-ids to complete checkout. Therefore, in order to launch the experiment, the Gross conversion should decrease, and the Retention should increase. Given that the screener should only affect the number of user-ids to complete checkout, but not the number of user-ids to remain enrolled past the 14-day boundary, the Net conversion should stay the same, in order to launch the experiment.

Measuring Standard Deviation

- **Gross conversion**: 0.0202
- **Retention**: 0.0549
- **Net conversion**: 0.0156

The unit of diversion for the experiment is cookies. It is equal to the unit of analysis in Gross conversion and Net conversion. However, the unit of analysis in Retention is the number of user-ids. Therefore, the analytical standard deviations for Gross conversion and Net conversion likely match the empirical standard deviation seen in the experiment, but the analytical standard deviation for Retention may not likely match the empirical standard deviation.

Sizing

Number of Samples vs. Power

Because the Gross conversion and Net conversion are highly correlated, Bonferroni correction is not used in the analysis. Using the rates from the baseline sample, with alpha = 0.05, and beta = 0.20, sample size for each evaluation metric is calculated using online calculator.

ratio of page views to clicks = 0.08
ratio of page views to enrolls = 0.0165

Number of page views for
- **Gross conversion**: 25,835/0.08 *2 = 645,875
- **Retention**: 39,115/0.0165 *2 = 4,741,213
- **Net conversion**: 27,413/0.08 *2 = 685,325

The largest page views is 4,741,213 for Retention.

Duration vs. Exposure

In this experiment, only an additional screen with message pops-up before the checkout and enrollment of the free trial, I do not think there is a chance that anyone gets hurt because of the duration of this experiment. In addition, no sensitive data is involved in this experiment. Therefore it is not risky for us to consider divert the entire traffic.

Given that not other experiments are performed at the same time, I would divert the entire traffic of Udacity for this experiment.

Number of days for

- **Gross conversion**: 645,875/40,000 = 17 days
- **Retention**: 4,741,213/40,000 = 119 days
- **Net conversion**: 685,325/40,000 = 18 days

Since it will take more than 100 days for Udacity to divert its entire traffic, the duration for Retention is too long. Therefore, Retention is excluded from the evaluation metrics. After iterating the evaluation metrics, the number of page views now is 685,325, while the duration is only 18 days for this experiment, which is practically feasible.

Experiment Analysis

Sanity Checks

- **Number of cookies**

The control group: 345,543
The experimental group: 344,660

Total number of page views: 690,203
Expected probability of a cookie into a control or experimental group: 0.5
Standard Error (SE): SQRT(0.5*(1-0.5)*(1/(345,543+ 344,660))) = 0.0006
Margin of error (m): SE * 1.96 = 0.0012
Confident Interval (CI): [0.5 -m, 0.5 +m] = [0.4988, 0.5012]

Observation: 345,543/ 690,203 = 0.5006

The Observation value is within the CI, thus it passed the sanity check.

- **Number of clicks**

The control group: 28,378
The experimental group: 28,325

Total number of clicks: 56,703
Expected probability of a click into a control or experimental group: 0.5
Standard Error (SE): SQRT(0.5*(1-0.5)*(1/(28,378 + 28,325))) = 0.0021
Margin of error (m): SE * 1.96 = 0.0041
Confident Interval (CI): [0.5 -m, 0.5 +m] = [0.4959, 0.5041]

Observation: 345,543/ 690,203 = 0.5005

The Observation value is within the CI, thus it passed the sanity check.

Result Analysis

Effect Size Tests

Only data within the period of October 11 and November 2 are taken into account, because of no data entry for Enrollment and Payment after November 2.

Bonferroni correction is not used because of the Gross conversion and Net conversion is highly correlated. Alpha = 0.05, Z-score = 1.96.

- **Gross conversion**

The control group: Clicks = 17,293
Enrollments = 3,785
Gross conversion = 0.2189

The experimental group: Clicks = 17,260
Enrollments = 3,423
Gross conversion = 0.1983

Difference = 0.1983 - 0.2189 = -0.0205
SE = 0.0044
m = 0.0086
CI = [-0.0291, -0.0119]

d_min = +/- 0.01
Statistically significant, since CI does not contain 0.
Practically significant, since CI does not contain the value of d_min.

- **Net conversion**

The control group: Clicks = 17,293
Payments = 2,033
Net conversion = 0.1178

The experimental group: Clicks = 17,260
Payments = 1,945
Net conversion = 0.1127

Difference = 0.1127 - 0.1178 = -0.0049

SE = 0.0034
m = 0.0067
CI = [-0.0116, 0.0019]

d_min = +/- 0.0075
Statistically NOT significant, since CI contains 0.
Practically NOT significant, since CI contains the value of d_min.

Sign Tests

A binomial sign test is performed to test the two evaluation metrics. Day-by-day data is used to evaluate whether the difference between experimental and control groups is positive or negative (experimental - control). If the difference is positive, we counted as a success event, vice versa the negative as a failure event.

- **Gross conversion**

Success: 4
Total days: 23
Probability for binomial test: 0.5
Two-tailed p-value: 0.0026

P-value < 0.05, indicating statistically significant of gross conversion.

- **Net conversion**

Success: 10
Total days: 23
Probability for binomial test: 0.5
Two-tailed p-value: 0.6776

P-value > 0.05, indicating not statistically significant of net conversion.

Summary

Bonferroni correction is not used in the analysis. The Gross conversion and Net conversion might be highly correlated. In order to launch the experiment, these two evaluation metrics need to be both matched our expectations (i.e. the Gross conversion should decrease and the net conversion should stay the same). When considering these two evaluation metrics together to make a decision, Bonferroni correction is not needed to adjust the alpha.

The effect size test indicated Gross conversion is both statistically and practically significant, while not in the Net conversion. The sign test also indicated a significant drop rate in the experimental group, which is aligned with the effect size test. There was no discrepancy between the hypothesis tests and the sign tests. The result suggested that the screener pops-up effectively decrease the number of students who enrolled at the initial clicks. On the other hand, the net conversion was not significantly affected by the change of screener pops-up.

Recommendation

The intension of this experiment was to “reduce the number of frustrated students who left the free trial because they didn't have enough time, but without significantly reducing the number of students to continue past the free trial and eventually complete the course”. The screener pops-up showed a significant effect on decreasing the Gross Conversion, reducing the number of students who attempted to enroll at the initial clicks. However, although there was not a statistically significant change in the Net Conversion, the confident interval (-0.0116) included the negative of the practical significance boundary (-0.0075). It indicates the number of students to remain enrolled past the 14-day boundary slightly dropped down, which would matter to the business. Considering to this potential risk, I would recommend not to launch this experiment.

Follow-Up Experiment

In order to increase the number of students to remain enrolled past the 14-day boundary, a follow-up experiment may be designed as followed:

After the 14-day free trial period, a feedback screen will pop-up with some suggestions to the student, including a summary of the time dedicated within the 14 days, estimated time needed for the rest of the course according to this pace, and suggestions on the pre-requisites skills for this course, to improve the learning experiences.

My hypothesis is that students who receive the feedback screen after the 14-day free trial period may re-evaluate their learning experience more objectively, so that they can make a more careful decision on the course cancellation. In the case, the number of students to remain enrolled past the 14-day boundary will increase.

Initial Unit of Diversion: the number of user-ids

Invariant Metric:

- **Number of user-ids**: number of user-ids to complete checkout and enroll in the free trial.

Evaluation Metric:

- **Retention**: 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.

Because the experiment will perform on enrolled students who already have stable user-ids, the user-ids will be chosen as the unit of diversion. The number of students who complete checkout should not change across the experimental and control groups, thus the number of user-ids is chosen as an invariant metric. Retention is chosen as the evaluation metric, because of its user-id base. If the hypothesis was true, retention should have an increase in the experimental group.

Since many students may not sure about whether their learning pace and their prerequisite knowledge got from the 14-day free trial learning is capable for the rest of the course or not, an objective feedback from Udacity will be necessary, to encourage them to stay in the program.