Design an A/B Test for Udacity Free Trial Screener

By Yuheng Cai

Project Overview

Udacity courses currently have two options on the home page: “start free trial” and “access course materials”. In the experiment, Udacity tested a change where student was prompted to input time commitment per week for a course, and then based on student input he/she was redirected to either enrolling in the free trial or access the course materials for free. Udacity could improve the overall student experience and improve coaches’ capacity to support students who are more likely to complete the course

This project is to design an A/B test for the actual experiment and to validate the hypothesis statistically.

Experiment Design

Metric Choice

Invariant metrics:

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

  Since it’s evenly distributed in control and experiment group, and it’s captured before student seeing the new change, number of cookies should not be impacted by the experiment and considered invariant.

- Number of clicks on “Start free trial” - number of unique cookies to click the “Start free trial”

  The click happens before student sees he experiment, so it’s independent from the experiment and considered invariant.

Evaluation metrics:

- Gross conversion - number of user-id to complete checkout and enroll in the free trial divided by number of unique cookies

  Since it’s directly reflecting the effect of the experiment where gross conversion in the control group would be expected to be higher than that in experiment group; therefore, this can be used as an evaluation metric for validating the hypothesis.

- Net conversion - number of user-ids to remain enrolled past the 14-day boundary divided by the number of unique cookies to click the ‘Start free trial’ button

  Can’t be used as invariant metric because students in experiment group will be prompted the time commitment upfront, which potentially leads to decrease of enrolment;

and for the exact same reason, it can be used as evaluation metric to measure the impact on the number of students continued in program after the free trial.
Net Conversion is preferred over Retention is due to the fact that Net Conversion uses number of cookies as denominator, which is also Unit of Diversion. Keeping Unit of Analysis and Unit of Diversion same will make analytical estimate comparable to the empirical variability.

Other metrics not considered:

- Number of user-ids - number of users who enroll in the free trial

  This metric is dependent on the experiment, we expect to see different value in the control group and experiment group, so it can't be an invariant; on the hand, number of user-id doesn't normalize the different sized experiment and control groups as Net Conversion does. So Net Conversion is a better choice for evaluation.

- Retention - number of user-ids to remain enrolled past the 14-day boundary divided by number of user-ids to complete checkout

  If students are explicitly prompted to take into account their time commitment, we should expect higher retention for those (in experiment group) choosing to enroll the course, compared with control group where students may not be aware of the time commitment and so dropout rate is expected to be higher. So this metric shouldn't be used as invariant metric.
  This could be an option for evaluation metric, but I prefer Net Conversion over Retention is due to the fact that Net Conversion uses number of cookies as denominator, which is also Unit of Diversion. Keeping Unit of Analysis and Unit of Diversion same will make analytical estimate comparable to the empirical variability.

- Click-through-probability on "Start free trial" - number of unique cookies to click "Start free trial" button divided by number of unique cookies to view the course overview page

  This could be an option for invariant metric because students in experiment group see the change only after clicking 'start free trial'. Having said that Number of Cookies and Number of clicks on "Start free trial" should have enough coverage. Adding this metric doesn't provide any new aspect of information or value. So it's not selected for invariant metric.

In order to make decision to launch the experiment, it's expected that:

- The Gross Conversion will decrease, which means the experiment will filter out some students without time commitment
- The Net Conversion will not decrease significantly so as to confirm the change won't introduce significant business impact while improving user experience.

Measuring Standard Deviation

Rough estimates of the baseline values is provided here, which is a modified version of Udacity's true numbers.

Gross Conversion when 3200 clicks & 40000 unique cookies to view pages:
    se = sqrt(0.20625*(1-0.20625)/3200) = 0.007152599
Gross Conversion when 5000 unique cookies to view pages:
    se = 0.00715 * sqrt(40000/5000) = 0.0202

Net Conversion when 3200 clicks & 40000 unique cookies to view pages:
    se = sqrt(0.1093125*(1-0.1093125)/3200) = 0.005515979
Net Conversion when 5000 unique cookies to view pages:
    se = 0.005515979*sqrt(40000/5000) = 0.0156

Both Gross Conversion and Net Conversion use number of cookies as denominator, which is also Unit of Diversion and Unit of Analysis. So analytical estimate would be comparable to the empirical one.

Sizing

Number of Samples vs. Power

Bonferroni correction was not used in analysis phase. This is due to the fact that metrics in the experiment have high correlation. Bonferroni correction will be too conservative.

Using Evan Miller, Samples is derived as following:

Probability of enrolling, given click:
    Baseline conversion rate: 20.625%, d_min: 1%
    Samples needed: 25835

Probability of payment, given click:
    Baseline conversion rate: 10.93125%, d_min: 0.75%
    Samples needed: 27413 (chosen)

Total pageviews = 27413/0.08*2 = 685325

Duration vs. Exposure

Since the experiment would not affect existing user flows except prompting student what level of time commitment needed before directing them to either "Start free trial" (needing time commitment) or free access to content (without needing commitment). So the overall risk is very low. In order to shorten time for experiment, more than half (70%) of the traffic will be directed to the experiment:
    Duration = 685325/(40000*70%) = 25 days

Experiment Analysis

Sanity Checks

The experiment data is provided here.
Invariant metrics:

- Number of cookies
  Total control group pageview: 345543
  Total experiment group pageview: 344660
  Total pageview: 690203
  Probability of cookie in control and experiment group: 0.5
  Standard error = sqrt(0.5*(1-0.5)*(1/345543+1/344660)) = 0.0006018
  Margin of error = 0.0006018*1.96 = 0.0011796
  Confidence Interval = [0.4988, 0.5012]
  Observed value = 34660/690203 = 0.5006

- Number of clicks:
  Total control group clicks: 28378
  Total experiment group clicks: 28325
  Total pageview: 56703
  Probability of cookie in control/experiment group: 0.5
  Standard error = sqrt(0.5*(1-0.5)*(1/28378+1/28325)) = 0.0021
  Margin of error = 0.0021*1.96 = 0.0041
  Confidence interval = [0.4959, 0.5041]
  Observed value = 28378/56703 = 0.50046

From above sanity check, both observed values are within confidence interval. So sanity check is passed for both invariant metrics.

Result Analysis

Effect Size Tests

Gross Conversion Analysis:

Control Group

- Clicks: 17293
- Enrolment: 3785
- Gross Conversion: 0.2188746892

Experiment Group

- Clicks: 17260
- Enrolment: 3423
- Gross Conversion: 0.1983198146

Standard error = 0.004371675385
Margin of error = 0.004371675385 * 1.96 = 0.00856848375

Pooled Probability = 0.2086
d_hat = - 0.02055
Confidence Interval = [-0.0291, -0.0120]

Gross Conversion Analysis Conclusion:

- Statistically significant (confidence interval doesn't contain zero)
- Practically significant (confidence interval doesn't contain d_min: 0.01)

Net Conversion Analysis:

Control Group

- Clicks: 17293
- Payment: 2033
- Net Conversion: 0.1175620193

Experiment Group

- Clicks: 17260
- Payment: 1945
- Net Conversion: 0.1126882966

Standard error = 0.003434133513
Margin of error = 0.003434133513 * 1.96 = 0.0067
Pooled Probability = 0.2086
d_hat = -0.0048737
Confidence interval = [-0.0116, 0.0019]

Net Conversion Analysis Conclusion:

- Not statistically significant (confidence contains zero)
- Not practically significant (upper bound of confidence interval < d_min: 0.0075; lower bound of confidence interval < -d_min: -0.0075)

Sign Tests

Gross Conversion:

- Number of success: 4
- Number of trials: 23
- Probability: 0.5
- Two-tailed p-value: 0.0026

Since p-value=0.0026 < alpha level (0.025), it's statistically significant.

Net Conversion:

- Number of success: 10
- Number of trials: 23
- Probability: 0.5
- Two-tailed p-value: 0.6776

Since p-value=0.6776 > alpha level (0.025), it's not statistically significant.

Summary

Bonferroni correction was not used in analysis phase. This is due to:

1. Metrics in the experiment have high correlation. Bonferroni correction will be too conservative.
2. Bonferroni correction is for controlling type I errors (false positive). However, for this experiment, we expect all metrics to be significant to launch the change where the risk of type II errors (false negatives) increases as the number of metrics increase. Bonferroni correction may be applied if we expect to launch the change when any of the metrics is significant.

From result analysis above, Gross Conversion experiment result is both statistically and practically significant; however, Net Conversion experiment result is both statistically and practically insignificant.

Recommendation

Gross Conversion is expected to be decreased when setting the expectation for the evaluation metrics. This is because students not meeting the time commitment will be discouraged to sign up for free trial.

Net Conversion is expected NOT to decrease to the level of hurting the business if it doesn't help increase paid users. Unfortunately, Net Conversion experiment result ended up being decreased both practically and statistically insignificantly. The confidence interval includes the negative number of the practical significance boundary, which means it's possible that Net Conversion would go down by an amount that would hurt the business; lastly, the upper bound of confidence interval doesn't meet the positive number of the practical significance boundary, which mean while it may increase the number of paid users, but the increase is not up to the min bar of practical boundary.

To sum up, my recommendation is NOT to launch the change.

Follow-Up Experiment

If the ultimate business goal is to let students be fully aware of courses' commitment and prerequisite and screen out those not satisfying these conditions in order to make best of the coach resources, then I'll like to propose running an experiment where:

1. Students will be prompted to input their time commitment (same UX as the Screener in this project). Students without time commitment is expected to be screened out.
2. For those students does proceed further, they will be asked to take a short competence quiz to check whether they meet the prerequisite for the course. For those don't meet, they will be redirected to access to free content (same as those without time commitment).

The rational behind this follow-up experiment is that apart from time commitment, students' competency level or prerequisite knowledge is also crucial for them to get through 14-day trial and eventually make payment to attempt completing the course.

Null hypothesis by creating such competency quiz, the Net Conversion will not decrease significantly hence hurting the business bottom line.

The Screener and Competence Quiz will be randomly assigned to a Control & Experiment Group. The whole courses for Control Group would not change and remained the same while courses for Experiment Group will have the time commitment prompt and competence quiz pages.

Unit of Diversion is cookies, users once clicked 'Start free trial' button will be tracked.

Invariant metrics include:

- Number of cookies - number of unique cookies to view the course overview page
- Number of clicks on "Start free trial" - number of unique cookies to click the "Start free trial"

Evaluation metric is Net Conversion, which validates whether rendering a "competence quiz" helps Udacity improve the overall student experience and improve coaches' capacity to support students who are more likely to complete the course.