Project 7: Design an A/B Test

Jason Cabral

Experiment Design

Metric Choice

For this experiment, the invariant metrics will be Number of Cookies, Number of Clicks and Click-Through-Probability. The evaluation metrics will be Gross Conversion and Net Conversion.

Number of Cookies (invariant) – The number of cookies viewing the course overview page would not be affected by the experimental change, so it makes a good invariant metric as it should remain the same in the control and experiment groups. For this reason, it also makes a poor evaluation metric.
Number of User-ids – The number of users who enroll in the free trial is expected to be different between the control and experiment groups, so it would not make a good invariant metric. Because enrollment occurs after the experiment, this could be used as an evaluation metric, but as it is not normalized, it has been excluded from the analysis.
Number of Clicks (invariant) – Because the experiment comes after the action of clicking the “Start Free Trial” button, the count of these clicks should be the same in the control and experiment groups. This metric therefor makes a good invariant metric and poor evaluation metric.
Click-through-probability (invariant) – This is simply a ratio of unique cookie clicks to unique cookies and is thus a good invariant metric and poor evaluation metric for the reasons stated previously.
Gross Conversion (evaluation) – This metric shows the number of users who enroll per unique cookie click. One of the goals of the experiment is to determine the effect on enrollment, so this is an important evaluation metric. If there is a significant effect, then the control and experiment groups would have different values for the metric, making this a poor invariant metric. Because the change is intended to illicit more caution among users, we should see a significantly lower enrollment per click in the experiment group.
Retention – This metric shows the number of users who remain enrolled through the free trial divided by the number of users who enrolled after clicking the “Start” button. This metric would be a good proxy for showing the change in the number of “dropouts”, or users who realized after enrolling that they would be unable to commit the time needed to successfully complete the program. However, the sample size required to properly evaluate this metric is quite large, meaning the experiment would have to run much longer than intended. For this reason, it will not be evaluated and because it should be different between control and experiment groups, it will not be used as an invariant metric either.
Net Conversion (evaluation) – Similar to Gross Conversion and Retention, this metric shows the number of users who remain enrolled past the free trial by the number of unique cookie clicks. The experimental group should produce different results than the control, so it is not invariant. Because this metric is related to Retention, it may help us get an idea of the long term effects of the experiment. If the hypothesis holds true, that the number of students who continue past the free trial will not be significantly reduced, then we should not see a significant decrease in this metric.

Measuring Standard Deviation

Gross Conversion Standard Deviation: 0.0202
Net Conversion Standard Deviation: 0.0156

Both Gross Conversion and Net Conversion share their unit of analysis with the unit of diversion, a cookie, so the variability of each metric should be relatively low. Additionally, each metric is a probability and should have a binomial distribution. Because of these two factors, the analytical estimates and empirical variability should be comparable, so we will use the former.

Sizing

Number of Samples vs. Power

In order to affirm our hypothesis, both of the selected evaluation metrics will need to meet our expectations. In the case where all metrics need to meet expectations, a single negative result, including a false one (type II error), would prevent a launch recommendation. The Bonferroni correction is designed to minimize the risk of false positives (type I error) by requiring more strict p-values. The risk of false positives is greater in situations where a positive result on only one metric would result in a launch recommendation, which is more likely among independent metrics. Because the selected metrics are correlated and both are required to meet expectations, the risk of type I error is relatively low and thus a Bonferroni correction will not be used during the analysis.

In order to properly evaluate the results of the experiment, we will need at least 685,325 page views.

Duration vs. Exposure

100% of the traffic will be diverted to either the experimental or control group. With 685,325 page views required and a daily average of 40,000 pageviews, the experiment will be run for 18 days.

The main risk in this experiment would be a failure in the code that prevents a user from creating an account, enrolling or viewing course materials. As these are critical to the user experience, this should be tested thoroughly before conducting the experiment. The data collected is not sensitive outside of user-ids, which can be considered personally identifiable. User-ids will be handled with the same security that user accounts are within the framework of the infrastructure, so including them in the experiment does not pose additional risk. Furthermore, the privacy policy of the website informs the user of how personally identifiable information, including user-ids, may be stored, used or distributed. Because of the low risk and insensitive data, there is no reason to limit the traffic included in the experiment.

Experiment Analysis

Sanity Checks

Number of Cookies Confidence Interval: (0.4988, 0.5012)
Number of Cookies Actual Observed Value: 0.5006
Number of Cookies PASSES Sanity Check

Number of Clicks on “Start Free Trial” Confidence Interval: (0.4959, 0.5041)
Number of Clicks on “Start Free Trial” Actual Observed Value: 0.5005
Number of Clicks on “Start Free Trial” PASSES Sanity Check

Number of Click-through-probability on “Start Free Trial” Confidence Interval: (-0.0013, 0.0013)
Number of Click-through-probability on “Start Free Trial” Actual Observed Value: 0.0001
Number of Click-through-probability on “Start Free Trial” PASSES Sanity Check

All observed values were within expected confidence intervals, so each metric passed the sanity check.

Result Analysis

Effect Size Tests

Gross Conversion Confidence Interval Around the Difference: (-0.0291, -0.0120)
The result for Gross Conversion is Statistically and Practically Significant

Net Conversion Confidence Interval Around the Difference: (-0.0116, 0.0019)
The result for Net Conversion is not Statistically or Practically Significant

Sign Tests

Gross Conversion Sign Test p-value: 0.0026
The result of the Sign Test for Gross Conversion is Statistically Significant

Net Conversion Sign Test p-value: 0.6776
The result of the Sign Test for Net Conversion is not Statistically Significant

Summary

As stated earlier, Gross Conversion and Net Conversion are correlated and both metrics must meet our expectation in order to recommend launching the change, meaning we are at risk for type II error. The Bonferroni correction is useful for minimizing type I error, so it was not used in this analysis. The effect size tests and sign tests both show that Gross Conversion produces a statistically significant result, while Net Conversion does not.

Recommendation

My recommendation is that the change not be implemented. The Gross Conversion metric shows that significantly less users are enrolling because of the suggestion that successful participants spend a minimum of 5 hours per week studying. This weeds out those who are less serious about the program, leading to higher student satisfaction and allows Udacity to dedicate more of it's time and resources to the most committed students. Net Conversion on the other hand did not show a statistical difference between the control and experimental groups. A 95% confidence interval did however include the negative practical significance boundary, indicating that there is a possibility that the change would in fact lead to a significant reduction in enrollment beyond the trial period.

While the change would have the positive effect of reducing initial enrollments, the potential that this would ultimately lead to lower long term enrollments and a reduction in revenue renders it such that I can not recommend that the change be implemented.

Follow-Up Experiment

One major reason for early cancellations may be due to differences in expectations between the student and the course instructor. One way to close the gap would be requiring all students to pass a more difficult quiz at the beginning of the course that would set expectations for the student about what working knowledge is required to successfully complete the course. These questions could test a student's abilities in coding, math or something else relevant to the course and may take many hours to complete. This would not only ensure that the student was prepared for the course academically, but would also help show the student that a time commitment is required as well.

Experiment Design

The experiment I would run would occur after users enroll in a course and would split those users into a control and experiment group. The control group would be given a standard introduction, discussing briefly the topics covered in the course and the prior knowledge that would be necessary to understand the framework of the course, before being directed to Lesson 1. The experimental group would be given the introduction but then directed to the quiz. Upon successful completion of the quiz, the user would be directed to Lesson 1.

My hypothesis is that a smaller proportion of those who enroll would make it to Lesson 1, but a higher proportion of those who start Lesson 1 would make it to the 50% mark of the course within the expected time-frame. As each section of the course has an expected time of completion and students are expected to commit a minimum number of hours per week, this time-frame should not be difficult to calculate.

Metrics

The invariant metric is:
Number of user-ids – The number of users who enroll in the free trial.

The evaluation metrics are:
Course-Start-Probability – The number of user-ids that start Lesson 1 divided by the number of user-ids that enroll in the course. (The control should be close to 1)
Course Retention – The number of user-ids that complete 50% of the course within an expected time-frame divided by the number of user-ids that start Lesson 1.

The unit of diversion will be a user-id as only those who are enrolled in the course will be tracked. This will also be the invariant metric and our sanity check, since it should be the same in both groups as the change will occur after enrollment.

Expectations

Given the hypothesis, we would expect Course-Start-Probability to be lower in the experimental group as those users less prepared for the course or those not willing to dedicate an appropriate amount of time would not complete the quiz or start Lesson 1. Similarly, Course Retention should be higher in the experimental group as those students who start Lesson 1 in this group should be less likely than students in the control group to drop out of the course early and more likely to dedicate the necessary amount of time.

Additional Considerations

If a student does not drop the course but also does not reach the 50% mark within the expected time-frame, then this can be considered a negative result for the experiment as it suggests that the student was not properly prepared for the rigors of the course. Also, because courses can take many months to complete, it will need to be determined if the benefits of this change would outweigh the costs of running a lengthy experiment.