**A/B Testing to Decide Whether a Warning of Expected Study Time Before Students Starting Free Udacity Course Would Decrease Their Withdrawing Halfway**

**Experiment Description**

At the time of this experiment, 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. This screenshot shows the experiment:

**A screenshot of a cell phone

Description automatically generated**

If this early warning did significantly reduce the number of students who give up halfway, 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 user­id from that point forward. The same user­id cannot enroll in the free trial twice. For users that do not enroll, their user­id is not tracked in the experiment, even if they were signed in when they visited the course overview page.

**Null Hypothesis:** This early warning would not significantly reduce the number of students withdrawing halfway

**Alternative Hypothesis:** This early warning would significantly reduce the number of students withdrawing halfway, without significantly reducing the number of students to continue past the free trial and eventually complete the course

**Experimental Design**

**Metric Choice**

***Invariant Metrics:*** Number of cookies, Number of clicks, Click-through-probability

***Evaluation Metrics:*** Gross conversion, Retention, Net conversion

**Invariant Metrics**

Invariant metrics are those which remain invariant throughout the experiment. One could expect a similar distribution of such metrics both on control and experiment side. In the given experiment, the invariant metrics are as follows –

**Number of cookies:** That is, number of unique cookies to view the course overview page. This is expected to be the same for control and experiment groups because the change occurs after the students click ‘start free trial’

**Number of clicks:** That is, number of unique cookies to click the "Start free trial" button (which happens before the free trial screener is trigger). At this point in the funnel the experience is the same for all users in control and experiment groups, and therefore they would not be significantly different

**Click-through-probability:** That is, number of unique cookies to click the "Start free trial" button divided by number of unique cookies to view the course overview page. Since number of cookies and number of clicks are invariants, click-through-probability should be roughly the same for control and experiment group

**Evaluation Metrics**

Evaluation Metrics are chosen based on which there are significant reduction in students enrolling in free course and significant increase in the ratio of students who remain after the free trial ends and complete the course. Each evaluation metric is associated with a minimum difference (dmin) that must be observed for consideration in the decision to launch the experiment. With this in mind, the following conditions must be satisfied -

**Gross conversion:** That is, 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 (dmin=0.01). This is expected to be significantly lowered for experiment group as we do expect to significantly reduce the students who enroll the free course using warning

**Retention:** That is, 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 (dmin=0.01). Since we do not want to significantly reduce the number of students to continue past the free trial and eventually complete the course and we do expect to significantly reduce the students who enroll the free course using warning, retention is expected to increase or at least roughly stay the same for experiment group

**Net conversion:** That is, number of user-ids to remain enrolled past the 14-day boundary (and thus make at least one payment) divided by the number of unique cookies to click the "Start free trial" button (dmin=0.0075). Because we want to reduce the number of students who withdraw halfway, we expect this metric to be significantly higher for experiment group

**Unused Metrics**

**Number of user-ids:** The number of users who enroll in the free trial (dmin=50). This metric is expected to be lower in the experiment group since we want to reduce the number of students who enroll the free course. However, the possible reduction can be related to both “frustrated” and “resolute” groups. So, this metric is not distinctive.

**Measuring Variability**

The standard error of the current metrics depends on the sample size and can be computed either analytically or empirically. For each metric selected as an evaluation metric, its standard deviation, or standard error, is estimated analytically given a sample size of 5000 cookies visiting the course overview page using the baseline values. The baseline values and calculations are in the [excel sheet](Final+Project+Baseline+Values.xlsx).

|  |  |
| --- | --- |
| **Evaluation Metric** | **Standard Deviation** |
| Gross conversion | 0.0202 |
| Retention | 0.0549 |
| Net Conversion | 0.0156 |

If we have the ratios from historic data, i.e. retrospective studies, we should have the standard deviation of the metric in the theoretical population, shouldn’t we? Moreover, the theoretical distribution may not hold, so it is better to gather data and use **bootstrapping** to estimate these standard errors empirically.

**Sizing**

**Sample Size**

Determination of the sample size is a fundamental step. Based on the distribution of the metric, the sample size is a function of some parameters such as alpha, i.e. significance level, beta, false negative probability or type II error, baseline conversion, practical significance and so. In this experiment, alpha=0.05, power=0.8 were used for each metric. The sample size is calculated each separately using this [online calculator](http://www.evanmiller.org/ab-testing/sample-size.html).

Since I have three evaluation metrics, adjustment to significant level might be done. **Bonferroni correction** is one approach to such adjustment, and there are more approaches to do so. However, in my case, I did not adjust, because these metrics are closely related to each other, and Bonferroni would be too conservative, causing inflation of the sample sizes.

**Pageviews for each evaluation metric to achieve target statistical power:**

***Gross conversion***

* Baseline Conversion: 20.625%
* Minimum Detectable Effect: 0.01
* Sample Size = 25,835 clicks/group
* Number of groups = 2 (experiment and control)
* Total sample size = 51,670 clicks
* Clicks/Pageview: 3200/40000 = 0.08 clicks/pageview
* Pageviews Required: 51670/0.08 = 645,875

***Retention***

* Baseline Conversion: 53%
* Minimum Detectable Effect: 0.01
* Sample Size = 39155 enrollments/group
* Number of groups = 2 (experiment and control)
* Total sample size = 78230 enrollments
* Enrollments/Pageview: 660/40000 = 0.0165 enrollments/pageview
* Pageviews Required: 78230/0.0165 = 4,741,212

***Net conversion***

* Baseline Conversion: 10.93125%
* Minimum Detectable Effect: 0.0075
* Sample Size = 27413 clicks/group
* Number of groups = 2 (experiment and control)
* Total sample size = 54826 clicks
* Clicks/Pageview: 3200/40000 = 0.08 clicks/pageview
* Pageviews Required: 54826/0.08 = 685,325

It’s clear that retention is the dominant metric and over 4 million total pageviews are needed for control and experiment groups. However, considering Udacity’s daily traffic is 40000 pageviews, 4 million total pageviews require more than 100 days to complete. It’s better not to guide 100% of the traffic to the experiment either, because the experiment may cause some unexpected side-effects. Therefore, the duration needs to be reduced. We can loosen the power of the test, and increase the alpha for the retention, since this metric is our bottleneck. Or increase the practical significance to 2%, instead of the default 1%. Then the total unique cookies required would be 1,185,455, and the minimum days would be reduced to 30.

Increasing the practical significance means that even though the change in the metric is statistically significant and practically significant based on the old criterion, now we deem it as insignificant. A good decision?  While it is worth trying, but the provided data about the traffic is insufficient for keeping the retention. So, we drop the retention, and continue with the net and gross conversion metrics.

Now, net conversion is dominant, and we need 35 days of duration of experiment supposing 50% of daily traffic is exposed to the experiment.

**Analysis**

The data from control and experiment groups is [here](Final+Project+Results.xlsx). This sheet also contains all the calculation steps for the each of the following checks and test. This data contains raw information needed to compute the above metrics, broken down day by day.

**Sanity Check**

The first step in validation of the experiment is assessment of invariant metrics, that is, checking whether my invariant metrics are equivalent between the two groups. With 95% confidence interval, the results are shown:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Lower bound | Upper bound | Observed | Pass |
| Number of cookies | 0.4988 | 0.5012 | 0.5006 | Yes |
| Number of clicks on ‘start free trial’ | 0.4959 | 0.5041 | 0.5005 | Yes |
| Click-through-probability on ‘start free trial’ | 0.0812 | 0.0830 | 0.0822 | Yes |

**Effect Size Tests**

For each evaluation metric, a 95% confidence interval is computed for the difference between control and experiment group. The result is statistically significant only when the 95% confidence interval does not include zero and practically significant only when the minimum detectable effect is not included. Although retention is not used as an evaluation metric, the confidence interval is also calculated.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | dmin | Lower bound | Upper bound | Statistical significance | Practical significance |
| Gross conversion | 0.01 | -0.0291 | -0.0120 | Yes | Yes |
| Retention | 0.01 | 0.0081 | 0.0541 | Yes | No |
| Net conversion | 0.0075 | -0.0116 | 0.0019 | No | No |

**Sign Tests**

**I ran sign test on each of my evaluation metrics using the day-by-day date. The sign test is just another method to validate the result obtained above, which gives the probability of X number of occurrences of one outcome given total two outcomes if the two outcome has equal probability of occurrence. Alpha of 0.05 is used.**

|  |  |  |
| --- | --- | --- |
|  | **p-value** | **Statistical significance** |
| **Gross conversion** | **0.0026** | **Yes** |
| **Retention** | **0.6776** | **No** |
| **Net conversion** | **0.6776** | **No** |

**.**

**Summary**

The requirement for launching the experiment is that the null hypothesis must be rejected for all evaluation metrics and that the difference between branches must meet or exceed the practical significance threshold. From sanity checks, we see that all our invariant metrics hold no significant differences between control and experiment groups. For our evaluation metrics, gross conversion is both statistically significant and practically significant, while net conversion is neither statistically significant nor practically significant. Sign tests agree with effect size tests in that gross conversion has significant different number of positive and negative signs for the differences between control and experiment groups, while the difference of positive and negative signs for net conversion is not significant.

**Recommendation**

**While it is certain that we expect a reduction in the number of students who enroll in this free trial by showing a warning before students enroll, specifically, we want the reduction to result from frustrated students who could not complete the course due to ending of the free trial instead of the students who could continue the course after free trial. Therefore, we do want gross conversion to be statistically and practically lower in experiment group. An insignificant difference in net conversion is exactly what we expect, because we do not want to reduce the number of students who would continue the course after free trial. In conclusion, I would launch this experiment. Before running the A/B test, I would first run an A/A test to validate the sensitivity, or robustness of my evaluation metrics – gross conversion and net conversion.**