A/B Testing to Determine an Effective Approach to Reduce Early Udacity Course Cancellation

Shahrzad Amoozegar

Introduction:

A/B testing refers to a randomized experimentation process in which two versions of a variable are compared against each other to determine which version drives business metrics. It is widely used for designing websites to make data-driven decisions based on statistical analysis. In this work, we'll perform an experiment on Udacity [1] website, an online learning platform, to drive data-backed decisions about the design of a web page.

Experiment Overview:

At the time of this experiment, students had two options to use Udacity courses: "start free trial", and "access course materials". Under "access course materials" they were able to audit courses for free, but they did not have access to coaching support, or verified certificate. By clicking "Start free trial", students were enrolled in a free trial for the paid version, and they were asked to enter their credit card information. After 14 days of learning for free, they were automatically charged unless they canceled the subscription.

In this experiment, Udacity tested a change where if the student clicked "start free trial", they were asked how much time they could devote to the course. If they indicated 5 or more hours per week, they were taken through the checkout process as usual. If they indicated fewer than 5 hours per week, a message was displayed indicting 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 had the option to continue enrolling in the free trial or access the course materials for free instead.

The first objective of this experiment was enhancing user experience by setting clear expectations prior to enrolment, and as a result reducing the number of frustrated students who couldn't allocate enough time to the program and left the program at the free trail.

The second objective was allocating coaching support to more dedicated students that were more likely to complete the program. The hypothesis for this experiment is designed as follows:

Null Hypothesis: Adding "free trial screener" might not be effective in reducing the early Udacity course cancellation.

Alternative Hypothesis: Adding "free trial screener" might reduce the number of frustrated students who left the free trial because they didn't have enough time without significantly reducing the number of students who past the free trial and eventually completed the course.

Experiment Design:

1.Unit of diversion

Unit of diversion indicates the subject of the test. One consideration in choosing unit of diversion is consistency of the test subject. For tests that are visible to users, it's necessary to use consistent unit of diversions such as cookies and user-ids. Since user-ids are not tracked for users that are not enrolled, in this experiment anonymous id-based unit of diversion(cookie) is used as the test subject.

2. Metric Choice

1.Gaurdrail or Invariant metrics:

Invariant metrics are used to ensure that there is consistency between control and experiment groups. In this experiment, Number of cookies, number of clicks and click through probability are used as invariant metrics.

a. **Number of cookies** indicates the number of unique cookies that visited the overview page of the website and is expected to have even distribution among control and experiment group.
b. **Number of clicks** indicates the number of unique cookies to click the "Start free trial" button. This metric is expected to have even distribution for control and experiment group.
c. **Click through probability** indicates the number of unique cookies who clicked "start free trial" divided by the number of unique cookies who viewed the overview page of the website . We expect this ratio to be equal for both control and experiment groups even though the number of cookies in the ratio might be slightly different.

2. Evaluation metrics:

Evaluation metrics are selected based on business requirements to ensure that business questions are answered. They capture differences between control and experiment group as a function of experiment.

a. **Gross conversion** is defined as the number of user-ids to complete checkout and enroll in the free trial divided by the number of unique cookies to click the "Start free trial" button. If the experiment group shows statistically significant decrease in Gross conversion rate, this new feature could lead to reducing the number of students who proceed to the free trial.
b. **Net conversion** is defined as the number of user-ids who remain enrolled beyond 14 days of free trial and thus make a payment, divided by the number of unique cookies who click the "Start free trial" button. This metric shows how many students remain enrolled after the free trial. If the experiment doesn't show significant decrease in net conversion, then we would be interested in launching the new feature.
c. **Retention** is defined as the number of user-ids to remain enrolled beyond 14 days of free trial and thus make a payment divided by the number of unique cookies who

complete the checkout. If this metric increases for the experiment group, we would be interested in launching the new feature.

We would launch the experiment if we observe statistically significant change in gross conversion and retention and no significant decrease in net conversion.

3. Duration and experiment size:

To determine the duration of a test, we need to consider several factors:

- **Statistical Power (1-*β*):**
  Statistical power is the probability of detecting a meaningful difference between different variants when there really is one. In an online website it means launching a feature when there is truly a positive improvement in business metrics. Statistical power holds an inverse relationship with Type II errors (β). The statistical power to perform this test is 80%.

- **Significant level (1-*α):***
  Significance level is the probability of committing the error of deciding that the statistical null hypothesis should be rejected, when in fact, one should have refrained from rejecting it. In an online website it means refraining from launching a feature when there is no positive improvement in business metrics. Statistical level holds an inverse relationship with Type I errors (α). The significant level to perform this test is 95%.

- **Minimum detectable effect:**
  Minimum detectable effect is the minimum improvement that business expects from implementing A/B test. In large companies like google even a 0.2% change is practically significant and would lead to large amounts of revenue increase. For this experiment, Udacity considers a minimum detectable effect of 0.01 to be practically significant.

- **Day-of-week effect:** The day-of-the-week effect relates to the user behavior that vary across days of the week. It is important to design a test that can capture the weekly cycle. The population of users might be different depending on the day of the week. As a result, a minimum of one week is required for running the test [4].

- **Novelty and Primacy effect:** When there is a new change in a product or feature, users might react differently some might be reluctant to change, and some might embrace it and start using the feature a lot. As a result, the test might have different outcomes at the beginning of the experiment, and it’s necessary to run the test long enough to ensure that the behavior of users get stabilized, and we reach plateau stage [4].

- **Target population**: Population depends on what subjects are relevant for the study under consideration. Is the experiment designed for a specific geographic location? Is it designed for a specific type of device? (Computer vs cell phone) Is the experiment designed for a certain demographic in the population? (For example, women between the age of 20-30). For testing this new feature for Udacity, we are interested to include all the population without any constraints on demographics, region, or device.

Sizing:

To estimate the size of the experiment, the analytical standard deviation and Baseline values are used to determine the required experiment size:

1. **Standard Deviation:**

Based on the definition of the hypothesis, we are interested in the difference between evaluation metrics for experiment and control group. Pooled standard error can be used to give a good comparison of both. Suppose x_c and x_e are number of users who click for control and experiment groups, respectively. Nc and Ne are the corresponding total number of users. The pooled probability is then,$\hat{P}$ =$\frac{xc+xe}{Nc+Ne}$ and pooled standard error of difference is:

$$s_p = \sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{Nc}+\frac{1}{Ne}\right)}$$

This formula is used to calculate the standard deviation of evaluation metrics based on the baseline values for each metric.

Probability of enrolling, given click	0.2062
Probability of payment, given enroll	0.53
Probability of payment, given click	0.1093

	Standard Deviation
Gross conversion	0.020231
Net conversion	0.054949
Retention	0.015602

Given the constraints of this experiment, and the value of standard deviations, we would not consider retention as an evaluation metric due to high variance and instability.

2. **Sample size:**

Minimum sample size required based on the power, statistical significance and practical significance is calculated using an online calculator.[2]

sample size for Gross conversion	25835
sample size for Net conversion	27413

3. Experiment size:

After calculating sample size, the total number of page views for each metric is determined based on baseline values reported by Udacity website.

Baseline values:

- Unique cookies to view course overview page per day: 40000
- Unique cookies to click "Start free trial" per day: 3200
- Enrollments per day: 660

Total number of pages required for each metric

Number of page views required for Gross conversion	645876
Number of page views required for Net conversion	685326

Duration of the experiment:

After determining the minimum number of page views for the experiment, the experiment duration would be calculated based on the number of unique cookies visited the overview page on a daily basis. To get minimum number of 635,326 page views, the experiment should run for at least 18 days, this would also prevent the potential day-of-week issue.

Other considerations for A/B testing:

Exposure:
One the design decisions for A/B testing is deciding which fraction of the traffic should be exposed to the experiment. Since this experiment is risk free and doesn’t seem to have physical, psychological, and emotional, social, and economic harm, we can expose 100% of traffic to get the required size for the experiment in a timely manner.

Safety:
One of the important considerations is the safety of the test. For large changes where companies are uncertain about how users might react, they should start with a smaller proportion of the users

first. The change we are considering for Udacity website is a risk-free change since the users' reactions is not unpredictable in this case, and it does not affect the privacy of users or doesn't involve sensitive data.

Spillover effect:
Another important aspect to consider when designing A/B testing is the effect of spillover. Spillover or interference between control and experiment group is more prevalent in social networks and two-sided markets. In these cases, users' behavior is likely impacted by that of people in their social circles or the shared resources between control and experiment group. Since Udacity increases the number of coaching supports based on the number users enrolled in the program, there will not be competition between users in control and experiment group [4].

4. Analyzing results:

Sanity Checks:

Sanity check in A/B testing refers to the use of methodologies to confirm that the experiment was run properly. It's necessary to ensure that the population of experiment and control groups are comparable. Also, to capture the true effect of the experiment, the invariant metrics should not change for experiment and control groups. Number of cookies, number of clicks and click through probability are three invariant metrics that we determined at the beginning of the test.
Since cookies were randomly assigned to control and experiment groups, the number of cookies within control group follows binomial distribution. The confidence interval for fraction of cookies in control group can be estimated based on confidence interval for normal distribution. This process is repeated for number of clicks and click through rate.

Invariant metric	Lower bound for confidence interval	upper bound for confidence interval
number of cookies	0.498820392	0.501179608
number of clicks	0.495884496	0.504115504
click through probability	0.001295679	-0.001295679

These values are compared against the observed values of invariant metrics during experiment and the results confirm that all invariant metrics pass sanity check.

Effect Size Tests:

In effect size test is used to evaluate the statistical difference between control and experiment, and is measured by using values of mean differences. Since cookies were randomly assigned to control and experiment groups, the number of cookies within control group follows binomial distribution, and due to the large sample size, the distribution can be approximated by normal distribution. As discussed in previous sections, gross conversion captures the ratio of clicks that lead to enrolment, and net conversion captures the ratio of clicks that lead to payment. The difference in gross conversion ratio between control and experiment group ($\hat{d}$), and net conversion ratio between control and experiment group ($\hat{d}$) is statistically tested and the confidence interval is constructed for $\hat{d}$:

Evaluation metric	Lower bound for confidence interval	upper bound for confidence interval	$\hat{d}$
Gross Conversion	-0.029123358	-0.011986392	-0.02055
Net Conversion	-0.011604624	0.001857179	-0.00487

Since the confidence interval for gross conversion does not include 0 (direction of change is clear) and $\hat{d}$ is greater than dmin = 0.01 we conclude that difference in Gross Conversion is both statistically and practically significant.

Since the confidence interval for net conversion includes 0 and $\hat{d}$ is less than dmin=0.0075 we conclude that difference in net Conversion is not neither statistically nor practically significant.

Sign Tests:

Sign test is a non-parametric test to evaluate the consistent differences between pairs of observations. The hypothesis for the sign test is as follows:

- H0: No difference in median of the signed differences.
- H1: Median of the signed differences is less than zero.

Since the test statistic is expected to follow a binomial distribution, the standard binomial test is used to calculate significance, and the normal approximation to the binomial distribution can be used for large sample sizes[3].
The p-values associated with sign tests on gross conversion and net conversion are as follows:

Evaluation metric	p-value
Gross Conversion	0.0026
Net Conversion	0.6776

Since the p-value is less than α level of 0.05 for gross conversion, we reject the hypothesis test, and conclude that this result is unlikely to come about by chance and the difference is statistically significant according to both the sign test and the effect size test.

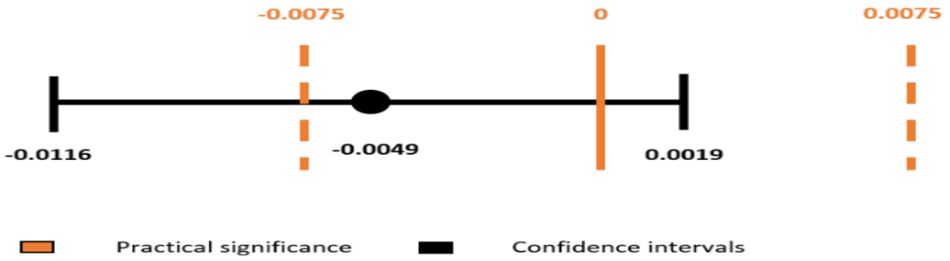
For net conversion, Since the p-value is less than α level of 0.05, we fail to reject the null hypothesis and conclude that the difference is not statistically significant.

5. Conclusion:

In this experiment we tested whether adding a new feature to Udacity website would have positive impact on the business, which is setting clearer expectations for students upfront, thus reducing the number of frustrated students who left the free trail due to lack of time. This test should be performed without significantly reducing the number of students to continue past the free trial.

Two evaluation metrics were defined for this experiment and the result of experiment showed statistically significant improvement in net conversion which is the ratio of clicks that lead to enrollment and no significance change in net conversion that were desired for this test.

For net conversion, the confidence interval does include the negative of the practical significance boundary. It's possible that this number went down by an amount that would matter to the business. In this case, we can run the experiment with more power (requires more samples) and this can shrink the confidence interval for reaching statistical significance. However, the better approach would be changing the intervention, since it is unlikely to change the overall trend in the net conversion metric.



Reference:

[1] https://www.udacity.com/

[2] https://www.evanmiller.org/ab-testing/sample-size.html

[3] https://en.wikipedia.org/wiki/Sign_test

[4] R. Kahovi, D. Tang, Y, Xu, Trustworthy Online Controlled Experiments : A Practical Guide to A/B Testing, 2020