Experiment Design

Stefan Dulman

Metric Choice

The invariant metrics are:

- Number of cookies: That is, number of unique cookies to view the course overview page. (dmin=3000)
- 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. (dmin=0.01)

The evaluation metrics are:

- 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)
- 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)
- 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)

Explanation on metric choices:

- Number of cookies: was chosen as invariant metric. This metric is a direct result of the marketing of the product should be roughly stable unless major advertisement campaigns take place. Cookies reach the website as a result of search engines, word-of-mouth, advertisements, etc. A small change in the content on such a large site should have almost no influence.
- Number of user-ids: This metric could have been used as an evaluation metric, linked to tracking the first part of the hypothesis. I prefered the other evaluation metrics due to the normalization factor they contain.
- Number of clicks: this was a candidate for an invariant metric. I chose to go with the next metric, as it felt a bit more robust.
- Click-through-probability: I chose this as an invariant as a certain percentage of both bots and humans (myself included) will most likely click a button and see what happens.

In my case I usually do it to estimate how complex the follow up process is and gather a bit more information.

- Gross conversion:by its definition, this metric is directly linked to the result of the experiment and was chosen as an evaluation metric.
- Retention: by its definition, this metric is directly linked to the result of the experiment and was chosen as an evaluation metric.
- Net conversion: by its definition, this metric is directly linked to the result of the experiment and was chosen as an evaluation metric.

Regarding the results we will be looking for, they can be split into several categories. First, the invariant metrics should stay constant. Then, the evaluation metrics should address the two parts of the hypothesis:

- Reduce the number of frustrated students leaving the free trial because of lack of time
- Without significantly reducing the number of students who make it past the trial and complete the course.

The first part of the hypothesis will be tracked by the gross conversion - which tells us how many students chose not to register in the free trial at all - after seeing the message presumably.
The second part of the hypothesis is tracked directly by the retention and net conversion. Both metrics provide a view on the numbers of students who become paying customers. Retention will probably provide a better view as its definition maps exactly to the second part of the hypothesis, while the net conversion includes also some students who just want to see how the registration process looks like (in the denominator).
In the case of a successful experiment, I would expect to see the gross conversion decreasing on the experiment group, while the retention and net conversion not decreasing (my expectation is that they will stay constant, although an increase will be welcome from the business perspective).

Measuring Standard Deviation

The 5000 pageviews correspond to 5000 * (3200 / 40000) = 400 cookies that click on “Start”.
The 5000 pageviews correspond to 5000 * (660 / 40000) = 82.5 cookies that enroll.

- Gross conversion standard deviation = sqrt(0.20625*(1-0.20625)/400) = 0.02023060414
- Retention standard deviation = sqrt(0.53*(1-0.53)/82.5) = 0.05494901218
- Net conversion standard deviation = sqrt(0.1093125*(1-0.1093125)/400) = 0.01560154458

Regarding the question of analytical versus empirical computation of variability, we need to look at the denominators of the metrics (the unit of analysis). For gross conversion and net conversion, the unit of analyses is cookies, which matches the unit of diversion. Hence, we can expect these two metrics to have similar analytical and empirical variabilities. In the case of retention, the unit of analysis is user id, thus it can be expected that the two variabilities will differ.

Sizing

Number of Samples vs. Power

Number of page views:

- Gross conversion - 645875
- Retention - 4741212
- Net conversion - 685325

Based on these results we will drop the retention metric as at 40000 page views per day it will take almost four months for the experiment to complete. For the remaining two metrics, we will need 685325 page views to obtain correct metric values.

Duration vs. Exposure

I would divert 100% of traffic for this experiment leading to a duration of approximately 18 days. The reasons for this choice are multifold:

- I consider the experiment not risky for the udacity. This is motivated by the fact that the user is presented with fair information, probably already made available by the company elsewhere.
- The duration of two and a half weeks is reasonable in terms of time with respect to the expected impact of the outcome of the experiment. At the end of the day, the experiment, if successful, will reduce the number of users who enroll/quit because of lack of time. This is not as powerful as an experiment that will increase the number of paying users, which should take priority.

Experiment Analysis

Sanity Checks

- Number of cookies: 95% confidence interval: [0.4988, 0.5012], observed value: 0.5006

- Clicks on “Start”: 95% confidence interval: [0.4959, 0.5041], observed value: 0.5005

Both metrics pass the sanity checks, as the observed values are inside the 95% confidence interval.

Result Analysis

Effect Size Tests

- Gross conversion - confidence interval: [-0.0291, -0.0119]
- Net conversion - confidence interval: [-0.0116, 0.00185]

The Gross conversion metric is both statistical and practical significant. The Net conversion metric is neither.

Sign Tests

- Gross conversion - p value - 0.0026
- Net conversion - p value - 0.6776

Only the Gross conversion is statistically significant.

Summary

Using the Bonferroni factor was not needed in this case as we were relying on both metrics at the same time to construct an answer to the hypothesis. In case we would have tested different variations (for example using segmentation of the populations), then the Bonferroni correction would have been useful.

There were no discrepancies between the sign test and the selected metrics.

Recommendation

Based on the metrics presented, I would recommend not introducing the change at this moment.

In lines with the goals of the experiment, to reduce the number of users who quit in frustration due to the lack of time, I think the experiment attained its purpose. The result in enrolling users is statistically significant. Unfortunately, the second metric is not statistically significant, meaning that the company might experience a decrease in revenue.

As only the first half of the hypothesis was verified, we cannot recommend introducing the change.

Follow-Up Experiment

Students may quit after the two weeks interval for a variety of reasons. I would like to target the subgroup of students who are not fully convinced yet that the course is beneficial for them because they did not have the time to evaluate it properly in the two weeks period. They would just notice the deadline of two weeks coming up and having their time allocated to some other activity in their life would decide to quit. As a possible solution, the Udacity team could offer an extension of, say, one more week at the end of the trial period. This additional week will be advertised only at the end of the first two weeks to the students who quit - no information about this extension should be presented elsewhere on the site.

The hypothesis would in this case be that the two weeks period is simply too small for a number of users, who then need to quit. In other words, we would like to see if adding another week of trial period once the two weeks ended would convince some of the students to stay.

The unit of diversion in this case is user id, as we are dealing only with registered users. As invariant metric we propose the number of user ids who register for the free trial.

As evaluation metrics we propose modified versions of the retention defined above:

- Retention_2_weeks - (the number of user ids who are paying customers after 2 weeks from registration) divided by (the number of user ids who finished the registration for the trial)
- Retention_3_weeks - (the number of user ids who are paying customers after 3 weeks from registration) divided by (the number of user ids who finished the registration for the trial)

In the case of a successful experiment confirming the hypothesis, I would expect the Retention_2_weeks not to decrease in the experimental group. The students who wanted to quit already quitted, and the ones who took the extra week option are not considered paying customers yet.

I would also expect the Retention_3_weeks to increase in the experimental group. This would show that some people were persuaded to continue and the metric will actually quantify the effect.