Evaluation of the Impact of Free Trial Screener: an A/B testing study

By Paul F. Seke E.

1. Background and objectives

Udacity performed an experiment whose objective was to increase the proportion of paying (and, eventually, graduating) students, i.e. students remaining enrolled beyond the free trial period (14 days) and, thus, performing at least one payment. In the experiment, students less likely to complete the course (as indicated during enrollment by self-reported inability to devote more than 5 hours per week to the course) were prompted to access the course materials for free instead. This change was expected to decrease the number of frustrated students living during free trial due to lack of time, without decreasing the number of students devoting enough time to the course. This way, should the experiment be successful, Udacity could improve the quality of the service to the students likely to complete the course, including the capacity of coaches to support them and the overall student experience. We analyzed the data obtained from the study as part of Udacity A/B testing course final project. Our objective was to come up with evidence-based recommendation whether to launch the change on Udacity website and, eventually, to suggest a high-level follow-on experiment.

2. Experiment Design

2.1. Metric Choice

2.1.1. Criteria for metric choice

The unit of diversion was a cookie. However, from enrollment in the free trial onward students were tracked by user-id. Three parameters whose marked difference between the control and experimental groups would have invalidated the experiment were selected as invariant metrics. On the other hand, three indicators of rates of student enrollment and retention at least up to first payment were used as evaluation metrics. The metrics available were:

- the number of unique cookies to view the course overview page, referred as "**number of cookies**" in the remaining part of this report (dmin, the practical significance boundary given as absolute change was 3000);
- Number of user-ids: That is, number of users who enroll in the free trial. (dmin=50);
- the number of unique cookies to click the "start free trial" button (which happens before the free trial screener is trigger) termed as "**number of clicks**" (dmin = 240);
- and the **click-through-probability** that is the ratio of the number of clicks to the number of cookies (dmin=0.01);
- the ratio of the number of users who enrolled in the free trial to the number of clicks termed as "**gross conversion**" (dmin= 0.01);
- the ratio of the number of enrolled users who made at least one payment to the number of clicks, termed as "**net conversion**" (dmin= 0.0075);

- and the ratio of the number of users who made at least one payment to the number of users who enrolled in the free trial, termed as "**retention**" (dmin=0.01).

The number of cookies and the number of clicks were selected as invariant metrics considering that marked differences between the control and experimental groups in this metric would have meant that cookie attribution to study groups is not random. This would have invalidated the experiment, as no valid comparison between these groups could have been possible. For similar reasons, the click-through probability (the ratio of these metrics, so mathematically not supposed to change as well) was also selected as invariant metric for the same reasons as the other evaluation metrics.

On the other hand, gross and net conversions and the retention were selected as evaluation metrics because they are indicators of changes in the number of users enrolled in the free trial and of students remaining enrolled beyond 14 days, the parameters supposedly affected by the experiment. The number of user_ids was not chosen as evaluation metric because it is an absolute measure, thus, it is highly susceptible to be altered by seasonal or weekly changes, hiding the real effects of the experiment. Unlike this metric, relative measures (ratios to the number of page views), such as net and gross conversions, are powerful evaluation metrics as they are poorly affected by seasonal or weekly changes in pageviews. The number of user_ids was not selected as invariant metric either, because user-ids being a reflect of the number of users enrolling, it was expected to go down as users less likely to graduate were discouraged to enroll.

2.1.2. Changes expected

The experiment was to be considered a success (with recommendation to launch the change) if all sanity checks were successful, and:

1) statistically significant decreases in gross conversion were observed in the experimental group (as result of decreases in enrolling users less likely to graduate), together with the absence of significant decrease in net conversion (this condition may also be reflected by an increased retention, as the denominator of this ratio is to decrease), and
2) the changes in each evaluation metric were meaningful for the business (confidence interval of change higher than practical significance boundary).

2.2. Measuring Standard Deviation

Given that for all the evaluation metrics we were dealing with binomial distributions with large numbers of unit of analysis, we assumed that the distributions were close to the normal distribution. Analytic standard deviations were considered good indicators of the empirical variability because the unit of diversion was the same type as the unit of analysis (metric denominators) for all the evaluation metrics. The analytic standard deviations of the evaluation metrics were: 0.0202 (gross conversion), 0.0156 (net conversion) and 0.0549 (retention).

2.3. Sizing

2.3.1. Number of Samples vs. Power

The pageview numbers required to power the experiment appropriately were the following: 631,375 (gross conversion), 685,325 (net conversion) and 4,741,212 (retention).

2.3.2. Duration vs. Exposure

It can be argued that the present is not risky as it neither exposes personal informations nor breaches ethical standards in use when dealing with personal informations. On this basis, 100% of the traffic could be devoted to the experiment. Collecting the number of pageview necessary to calculate the retention rate would require a long running time though (17 weeks). Considering that the information provided by the retention rate (number of enrolled and of paying students) is also provided by the gross and net conversions taken together, we did not use the first evaluation metric. This approach resulted in reasonable running time: only 18 days (with 47.59% page visitors seeing the change).

However, in marketing (the "company image" business), the potential impact (positive or negative) of visible website changes on a business reputation cannot be accurately anticipated. For instance, a change may harass visitors, be considered offending or some popular blogger may built a successful "conspiracy theory" about (it can be argued for instance that with the policy of encouraging users not to take the course if they do not have at least 5 hours per week Udacity and related companies are restricting the categories of people who can enroll, excluding workers and favoring unemployed users, or some more creative theory detrimental for the company image). This may give a bad reputation to the company, requiring hard marketing interventions like rebranding, whose pittfalls can be costly (http://uk.businessesforsale.com/uk/search/businesses-for-sale/articles/the-pitfalls-of-rebranding). So, considering that such risk for company image can be easily prevented by using a small fraction of the traffic during experiments, to our opinion it would be wise not to use 100% of the traffic. For instance, using 57.11% of the traffic (with only 28.55% of page visitors seeing the change), pageviews could be collected in 30 days.

3. Experiment Analysis

3.1. Sanity Checks

All invariant metrics passed the sanity check aimed at verifying their equivalence between the control and experimental group. More specifically, we assessed that invariant metric actual values were in the 95% confidence interval of the values we expected to observe. The confidence intervals (respectively the actual observed values) of the invariant metrics were:

- [0.4988; 0.5012] (0.5006) for the **number of cookies**
- [0.4959; 0.5041] (0.5005) for the **number of clicks**
- and [-0.0013; 0.0013] (0.0001) for the **click-through-probability.**

3.2. Result Analysis

3.2.1. Effect Size Tests

The 95% confidence interval around the difference between the experimental and control groups for the **gross conversion** was [-0.0291; -0.012]. The change was statistically significant, as the interval did not include 0. In addition, the change also had a practical significance, as the confidence interval did not include the practical significance boundary (dmin = 0.01). Instead, although the 95% confidence interval around the difference between the experimental and control groups for the **net conversion** ([-0.0116; 0.0019]) did not achieve a practical significance (dmin negative value -0.0075 being in the interval), the change tested in the experiment may have induced a reduction in students enrolled beyond

14 days relevant for the business as suggested by the fact that the lower boundary of the interval was negative.

3.2.2. Sign Tests

Sign tests assessing gross and net conversion successes were performed using day-by-day data. The tests revealed significantly fewer successes in gross conversion ($p = 0.0026$), but no significant change in net conversion ($p = 0.6776$). These results are in agreement with the effect size test observations (confidence interval for the difference).

3.2.3. Summary

Our analysis on the present Udacity A/B testing experiment used the number of unique cookies (**number of cookies**), the number of unique cookies to click the "start free trial" button (**number of clicks**) and the ratio of these numbers (**click-through-probability**) as invariant metrics; while the two evaluation metrics finally used were ratios of either the number of users enrolling in the free trial (**gross conversion**) or the number of users who made at least one payment (**net conversion**) to the number of clicks.

All invariant metrics passed the sanity checks. The number of samples necessary to confer the appropriate power to the effect size tests was determined without using Bonferroni correction, as we expected both our metrics to match our expectations in order to recommend the launch of the change.

As expected, the effect size test revealed a significant and practical decrease in gross conversion, accompanied by significantly fewer successes as revealed by the sign test. Still as expected, the sign test and the effect size test did not reveal any statistically or practically significant change in net conversion. However, the lower boundary of the 95% confidence interval around the difference between the experimental and control groups in the effect size test was lower negative, indicating that during the experiment a decrease relevant for the business was observed in the number of students remaining enrolled beyond 14 days.

3.3. Recommendation

Given that the experiment was accompanied by a decrease in the number of students remaining enrolled after 14 days, the change should not be launched.

4. Follow-Up Experiment

In the present experiment, Udacity aim was to “improve the overall student experience and improve coaches' capacity to support students likely to complete the course”. Instead, the next step from here could be an attempt to increase the number of users enrolling, remaining enrolled beyond 14 days and, eventually, completing the courses.

4.1. Theoretical considerations

4.1.1. Who is interested in data science?

To become data analyst/scientist, people would basically need 3 groups of skills (http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram):

(*i*) hacking skills (computer scientists more skilled here);
(*ii*) mathematics and statistics knowledge (statisticians more skilled here); and
(*iii*) substantive expertise (traditional researchers in non computer science fields, such as medicine, biology, economy, etc. more skilled here).

Therefore, people interested in data science (potential candidates for Data Analyst Nanodegree) are expected to have at least 2 of the 3 required skills, but will rarely have the 3 (hypothetically).

4.1.2. Limitations of attention while learning

To our opinion as teacher and neuroscientist, people learn by association of what is new to what is known/familiar (in a lesser extent, also according to their sensitivity as shown by the success of the 4MAT teaching method. Details here http://www.4mat.eu). So, people are more likely to learn when the new information can clearly be associated to the previous knowledge. Such a targeted approach allows an optimal use of student maximal attention time (the brain would devote only 20-45 min of effort when trying to learn. A good discussion about here: http://naturalhealthcare.ca/glossaries.phtml?term=sustained+attention#.V9PZTtEvC1M).

4.1.3. Our working hypothesis

On this basis, we propose that adapting the lectures to experts of each of the 3 categories with basic/intermediate understanding of the others may improve user self-conviction on the proper skills, increasing their motivation to learn. This is expected to result in a drastic decrease in the number of enrolled students dropping the courses, while increasing the number of users enrolling and completing the courses.

4.1.4. Examples of personalized curricula

Such “personalized” courses may consist, still in the data analyst nanodegree program example, for instance:

- *For people with extensive knowledge of programming in python but basic knowledge in statistics:* making Python and Pandas lecture optional may allow starting statistics lecture from high school level, with more basic statistics concepts (and possibly with appealing biostatistics or econometry quizzes and projects which may introduce them to other fields where they may shift to one day as data analysts/scientists)

- *For people with extensive experience in statistics (including traditional researchers) and basic knowledge of programming:* making descriptive and inferential statistics lecture optional may allow adding the Introduction to Computer Science to the programme, so they can strengthen their coding abilities in python language

- *Finally, people with strong skills in only one one of the three skills required and no confidence in their abilities in the others*, it should be suggested to start by taking basic to intermediate courses in the missing fields, as confidence in own skills can be critical for grasping better what is already known and acquiring new skills

4.2. Proposed study

4.2.1. Objective:

The study objective will be to assess whether background-dependent presentation of the course contents may increase the number of students enrolling and staying after 14 days (thus, eventually, graduating).

4.2.2. Experimental approach

1) The statement “Udacity offers a Data Analyst Nanodegree program tailored to your background” will be added above the area containing the buttons "start free trial" or "access course materials".
2) Once clicking on the "start free trial" button, users will be prompted to select 3 or 5 background skills in a list with 6 or 9 elements (with 2 or 3 elements of hacking, statistics and basic research skills).
3) Then, users enrolling will be presented with a course syllabus based on the main and secondary skills indicated (determined on the basis of the number of hacking, statistics or basic research elements selected).

4.2.3. Metrics

As the previous experiment, the unit of diversion will be a cookie, but student enrolled in the free trial will be tracked by user-id from that point onward.

Invariant metric

- number of unique cookies to view the course overview page (“**number of cookies**”);

Evaluation metrics

- **click-through-probability** that is the ratio of the number of unique cookies to click the "start free trial" button (“**number of clicks**”) to the number of cookies
- the ratio of the number of users who enrolled in the free trial to the number of clicks (“**gross conversion**”)
- the ratio of the number of enrolled users who made at least one payment to the number of clicks (“**net conversion**”).