A/B Test

Ryosuke Honda

August 31, 2016

1 Experiment Overview: Free Trial Screener

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. The hypothesis was that this might set clearer expectations for students upfront, thus reducing the number of frustrated students who left the free trial because they didn't have enough time— without significantly reducing the number of students to continue past the free trial and eventually complete the course. If this hypothesis held true, 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 user-id from that point forward. The same user-id cannot enroll in the free trial twice. For users that do not enroll, their user-id is not tracked in the experiment, even if they were signed in when they visited the course overview page.

2 Experiment Design

2.1 Metric Choice

List which metrics you will use as invariant metrics and evaluation metrics here. (These should be the same metrics you chose in the "Choosing Invariant Metrics" and "Choosing Evaluation Metrics" quizzes.)

Invariant Metrics

- Number of cookies
- Number of clicks

Since Udacity wants to improve the overall student experience and asks student whether they can devote their time more than 5 hours per week or not. This process is done before they enroll Udacity, so counting "Number of user-ids" which is crated "after" the enrollment is not effective. Students are asked questions when they click the "start free trial", so "number of clicks" is inevitable for invariant metrics. For the comparison, "number of cookies" is needed because if there's only one invariant metric, we can't decide whether the change is from the questions students are asked or not. Number of cookies and Number of clicks can be evenly divided in the control and experiment group in this experiment. Therefore I chose them as invariant metrics. On the other hand, I don't choose Number of user-ids since the number will be different between control and experiment group during this test.

Evaluation Metrics

- Gross conversion
- Retention
- Net conversion

I don't choose number of user ID as evaluation metric. If the hypothesis hold true, the number of user-id will decrease in the experiment group. However, smaller number of user-id in experiment group doesn't always mean the test is effective. Smaller number of user-id may be because of smaller number of pageviews of Udacity. So by checking only this metric, we can't know the cause of decrease. I don't choose Click-through-probability since if I chose it as an evaluation metric, I can't find the change between control and experiment group. Higher number of clicks doesn't mean less number of frustrated students. This metric is not relevant for this test.

Gross conversion 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. When students click the "Start free trial", they are asked questions whether they can devote enough time for Udacity. If they can devote enough time, they are encouraged to enroll. However if they can't, they are encouraged to access course materials. This will be effective to reduce the number of frustrated students since students who can't devote more than 5 hours per week will not enroll. If the hypothesis hold true, the number of students who try free trial will reduce. That means that gross conversion will reduce.Therefore, this metric is necessary.

Retention 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. Udacity doesn't want to reduce the number of paid students significantly by this test. This metric is also necessary because we can check whether the number of paid students are decreasing or not by this test.

If the hypothesis is true, the number of students who try free trial reduces without reducing the number of students who are enrolled past 14 days. This

means that retention will increase. Net conversion 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. This metric is also necessary because of the same reason described in Retention. If the hypothesis hold true, the number of students who remain enrolled past 14 days won't change so much. Therefore, this metric also doesn't change.

For each metric, explain both why you did or did not use it as an invariant metric and why you did or did not use it as an evaluation metric. Also, state what results you will look for in your evaluation metrics in order to launch the experiment.

3 Measuring Standard Deviation

List the standard deviation of each of your evaluation metrics. (These should be the answers from the "Calculating standard deviation" quiz.)

Gross conversion: 0.0202
Retention: 0.0549
Net conversion: 0.0156

As for Gross conversion and Net conversion, unit of diversion is equal to unit of analysis. Therefore, the analytical estimate would be comparable to the empirical variability.

As for Retention, unit of diversion is not equal to unit of analysis. Hence the empirical variability may be different from the analytical estimate. Therefore we perform both an analytical and empirical estimate for retention metric.

For each of your evaluation metrics, indicate whether you think the analytic estimate would be comparable to the empirical variability, or whether you expect them to be different (in which case it might be worth doing an empirical estimate if there is time). Briefly give your reasoning in each case.

4 Sizing

4.1 Number of Smaples vs. Power

Indicate whether you will use the Bonferroni correction during your analysis phase, and give the number of pageviews you will need to power you experiment appropriately. (These should be the answers from the "Calculating Number of Pageviews" quiz.)

I don't use Bonferroni correction during analysis phase. So the $\alpha = 0.05$ I selected Gross conversion, Retention, Net conversion as an evaluation metrics. The pageviews of Gross conversion, Retention, Net conversion are 645875, 4741212 and 685325 respectively. Therefore, to test the experiment, we need 4741212 pageviews.

4.2 Duration vs. Explosure

Indicate what fraction of traffic you would divert to this experiment and, given this, how many days you would need to run the experiment. (These should be the answers from the "Choosing Duration and Exposure" quiz.)

Unique cookies per day are 40000. It means it will take 4741212/40000=119 days to finish the experiment. This number isn't feasible. so I will not use Retention as an evaluation metric. Therefore the number of Sample will also change because of the change in bonferroni correction.

I use Gross conversion and Net conversion as evaluation metrics. In this case, I chose 685325(Net conversion) as a necessary pageview. I set the fraction of traffic as 1 since the collected data during this test isn't sensitive and isn't identifiable because the unit of diversion is cookie. In this case, the length of experiment is 685325/40000=19.7875 18. 18 days is feasible duration so I chose it as the length of experiment.

5 Experiment Analysis

5.1 Number of Smaples vs. Power

For each of your invariant metrics, give the 95% confidence interval for the value you expect to observe, the actual observed value, and whether the metric passes your sanity check. (These should be the answers from the "Sanity Checks" quiz.)

Number of cookies The number of cookies in control group and experiment group is 345,543,344,660 respectively. The total number of cookies is 690,203.

Standard Error(SE)=$\sqrt{0.5 \times 0.5 \times (1/690203)}$
$m = SE \times 1.96$
Lower bound=$0.5 - m = 0.4988$
Upper bound=$0.5 + m = 0.5012$
Observed value =$345543/690203 = 0.5006$

Lower Bound	Upper Bound	Observed
04988	0.5012	0.5006

Observed value is in between lower bound and upper bound. Therefore, number of cookies passes sanity check.

Number of clicks on "Start free trial" The number of clicks in control group and experiment group is 28,378,28,325 respectively. The total number of cookies is 56,703.

Standard Error (SE)=$\sqrt{0.5 \times 0.5 \times (1/56703)}$
m=$SE \times 1.96$
Lower bound=$0.5 - m = 0.4959$
Upper bound=$0.5 + m = 0.5041$
Observed value =$28378/56703 = 0.5005$

Lower Bound	Upper Bound	Observed
04959	0.5041	0.5005

Observed value is in between lower bound and upper bound. Therefore, number of clicks on "Start free trial" passes sanity check.

For any sanity check that did not pass, explain your best guess as to what went wrong based on the day-by-day data. Do not proceed to the rest of the analysis unless all sanity checks pass.

6 Result Analysis

6.1 Effect Size Tests

For each of your evaluation metrics, give a 95% confidence interval around the difference between the experiment and control groups. Indicate whether each metric is statistically and practically significant. (These should be the answers from the "Effect Size Tests" quiz.)

I used Bonferroni correction to calculate the effect test size. Therefore α =0.05/2 (I used two metrics: gross conversion and net conversion.)

$P_{pool} = \frac{X_{cont}+X_{exp}}{N_{cont}+N_{exp}}$
$SE_{pool} = \sqrt{P_{pool} \times (1 - P_{pool}) \times (1/N_{cont} + 1/N_{exp})}$
$\hat{d} = X_{exp}/N_{exp} - X_{cont}/N_{cont}$
$m = 1.96 \times SE_{pool}$
Lower Bound=$\hat{d} - m$
Upper Bound=$\hat{d} + m$

Gross conversion
$X_{cont} = 3785, X_{exp} = 3423, N_{cont} = 17293, N_{exp} = 17260$

Lower Bound	Upper Bound
-0.0291	-0.0120

Since the range between lower bound and upper bound doesn't include 0, it means that Gross conversion's effect test size is statistically significant. d_{min} for the gross conversion is 0.01. Both lower bound and upper bound are less than 0.01. This means that the effect size is practically significant.

Net conversion
$X_{cont} = 2033, X_{exp} = 1945, N_{cont} = 17293, N_{exp} = 17260$

Lower Bound	Upper Bound
-0.0116	0.0019

The range between lower bound and upper bound includes 0.Therefore net conversion's effect test size isn't statistically significant. d_{min} for the gross conversion is 0.0075. d_{min} exists between lower bound and upper bound. This means that the effect size isn't practically significant.

6.2 Sign Tests

For each of your evaluation metrics, do a sign test using the day-by-day data, and report the p- value of the sign test and whether the result is statistically significant. (These should be the answers from the "Sign Tests" quiz.)

Gross conversion:
Number of success:4
Number of trials:23
Probability:0.5
p-value is 0.0026
p-value is less than 5%,so gross conversion is statistically significant.

Net conversion:
Number of success:10
Number of trials:23
Probability:0.5
p-value is 0.6776
p-value is far more than 5%, so net conversion isn't statistically significant.

6.3 Summary

State whether you used the Bonferroni correction, and explain why or why not. If there are any discrepancies between the effect size hypothesis tests and the sign tests, describe the discrepancy and why you think it arose.

We have multiple metrics and we need all of them to meet criteria in order to launch in this A/B test. In this A/B test, we want to check the "DIFFERENCE" between control and experiment group. Therefore, if we can reject the null hypothesis, we can see the difference. Bonferroni correction is designed to decrease the risk of Type Ierror(null hypothesis is true, but rejected). This is not our purpose. Thus I don't use Bonferroni correction.

When it comes to Gross conversion, both effect size hypothesis and the sign test show statistically significant. When it comes to Net conversion, neither effect size hypothesis and sign test shows statistically significant, meaning there's no discrepancies between them.

7 Recommendation

7.1 Summary

Make a recommendation and briefly describe your reasoning.

I don't recommend the update.

According to Gross conversion, the number ratio of entering free trial has decreased and the number is statistically significant from the both effect size test and sign test. The decreased number means that the number of students who think they can't dedicate their time for more than 5 hours has decreased.This is good for Udacity because Udacity can improve coaches' capacity to support

students who are likely to complete the course.
For the net conversion, neither effect size test nor sign test are statistically significant. Also, the effect size test isn't practically significant. However, the confidence interval does include the negative practical significance boundary. This means that it is reasonably possible the number of students will decrease. Since Udacity doesn't want to lose students, I don't recommend to updata.

8 Follow-Up Experiment

8.1 Summary

Give a high-level description of the follow up experiment you would run, what your hypothesis would be, what metrics you would want to measure, what your unit of diversion would be, and your reasoning for these choices.

I want to suggest to set up free preliminary NanoDegree course. People who enter "free trial" have to finish pre-Nanodegree within 2 weeks. If they can't finish it, they are advised to access course materials(free). People who finish the pre-Nanodegree have a choice to enter the "Paid" nanodegree course. That is for everybody who wants to take nanodegree. This is because people who "THINK" they can devote their time for more than 5 hours per week may not actually spend time for more than 5 hours. Threfore, by setting pre-Nanodegree course, those people will realize that they can't devote 5 hours per week.

The hypothesis is this is that Udacity will enhance the number of motivated people who finish nanodegree and will help reduce the number of frustrated students and students who "thought" they can devote more than 5 hours. If this hypothesis hold true, Udacity can student experience and improve coaches' capacity to support students who are likely to improve the overallcomplete the course.

The unit of diversion is the user-id since the target is those who enter free trial.

Invariant metrics

- Number of user-ids: My purpose of this test is to check whether or not the test will reduce the number of frustrated students who have already enrolled. So, tracking the number of user-ids is significant.
- Number of people who pay at least once: My test is to reduce the frustrated students and students who "thought" they can devote more than 5 hours. In the experiment group, students have to finish the pre-Nanodegree within 2weeks and then, they are encouraged to proceed to take nanodegree(paid one). So, tracking number of people who pay at least once is also significant metric.

Evaluation metric

- Retention:number of user-ids to remain enrolled past the 14-day boundary (and thus make at least one payment) divided by the number of user-ids.

Those who finished the pre-Nanodegree realize that they are appropriate for taking nanodegree or not. So this ratio will change by setting pre-nanodegree. If the ratio reduced in the experimental group compared to control group, that means that the number of people who aren't meet the prerequisite is decreased. Therefore, this metric is necessary.