A/B Testing

Metric Choice

List which metrics you will use as invariant metrics and evaluation metrics here.

Number of cookies: Good invariant

Because the number of visits happen before the user sees the experiment, and are thus independent from it.As the *start free trial* page changes, the number of user that visit the website is unlikely to vary as that page has not been seen yet and should not affect users visiting the page.

Number of Clicks:invariance metric

The number of users (tracked as unique cookies at this stage) to click the free trial button. This is appropriate as an invariant metric but not an evaluation metric. Equal distribution amongst the experiment and control groups would be expected since at this point in the funnel the experience is the same for all users and therefore elements of the experiment would not be expected to impact clicking the "start free trial" button.

Click-through-probability: Invariance metric

Good invariant metric because the clicks happen before the user sees the experiment, and are thus independent from it.

Number of user-ids:

Since the enrollment depends on the rendering of *start free trial* page, I would expect to see discrepancies in the control and experimental group. As such, it cannot be an invariance metric. On the other hand, it makes for a poor evaluation metric as it is redundant compared to the other metrics. The number of user-ids or enrolled users can fluctuate a lot with respect to the number of *start free trial* clicks on a given day, and thus not a good proxy for this experiment. Instead, the number of user-ids divided by the number of *start free trial* clicks, which is the gross conversion, is a better metric as it marginalizes variances in the empirical count of user-ids.That is, number of users who enroll in the free trial. (dmin=50)

Evaluation Metrics: gross conversion, retention, net conversion

- Retention: Not a good invariant metric because the number of users who enroll in the free trial is dependent on the experiment. Good evaluation metric because it is directly dependent on the effect of the experiment, and also shows positive financial outcome of the change ,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 ,could measure whether or not the screener had an effect on the 14-day dropout rate
- Net Conversion: evaluation metric. The net conversion is the product of two evaluation metrics: gross conversion and retention, and it can be considered as a more general goal of the A/B test whether rendering a "5 or more hours per week" suggestion helps increase the ratio of users who make payment over those who see the start free trial page. Therefore it is also a good evaluation metric like rest of them .
- Gross Conversion: This is the number of user-ids to complete checkout and enroll in the free trial per unique cookie to click the "start free trial" button. dmin = 0.01,could measure whether or not the screener had an effect on enrollment.

To Launch the experiment the gross conversion will decrease practically significance, which indicate whether the cost will be lower by introducing the new screener; while net conversion will not decrease statistically significance, which indicate the screener whether or not affect the revenues.

Measuring Standard Deviation

List the standard deviation of each of your evaluation metrics.

Net conversion: 0.01560

Gros conversion: 0.0202

Retention:.05490

For each of your evaluation metrics, indicate whether you think the analytic estimate would be comparable to the 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.

Net conversion and Gross conversion both have the number of cookies as their denominator, which is also our unit of diversion. We can therefore proceed using an analytical estimate of the variance.

For Retention, the denominator is 'Number of users enrolled the courseware' which is not similar as Unit of Diversion. The unit of analysis and the unit of diversion are not the same therefore the analytical and the empirical estimates are different.The largest sample size is our limiting factor retention rate, so we require a total of pageviews to conduct the experiment.

Sizing
Number of Samples 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

I did not use the Bonferroni correction.

Pageviews for Each Evaluation Metric to Achieve Target Statistical Power

Gross Conversion

- Baseline Conversion: 20.625%
- alpha: 5%
- beta: 20%
- 1 - beta: 80%
- sample size = 25,835 per group
- total sample size = 51,670
- pageview: 3200/40000 = .08 pageview
- pageviews = 645,875

Net Conversion

- Baseline Conversion: 10.9313%
- alpha: 5%
- beta: 20%
- 1 - beta: 80%

- sample size = 27,413 per group
- total sample size = 54,826
- pageview: 3200/40000 = .08 pageview
- pageviews = 685,325

Retention:

- Baseline Conversion: 53%
- alpha: 5%
- beta: 20%
- 1 - beta: 80%
- sample size = 39,115
- pageview: 3200/40000 = .08 pageview
- pageviews = 4,741,212

Calculated using below .

- http://www.evanmiller.org/ab-testing/sample-size.html#!16.5;80;5;1;1
- http://graphpad.com/quickcalcs/binomial1.cfm

To achieve the target number of pageviews for the retention metric, it would take 117 days of complete site traffic, which is too long for an A/B test. Thus, only gross conversion and net conversion are used as evaluation metrics with the required number of pageviews being 685,325 and taking only 18 days at a 100% site traffic percentage.

Duration vs. Exposure

we would choose Gross conversion and Net conversion as evaluation metrics, and we abandon Net conversion as evaluation metrics since this would require too long duration.The fraction of Udacity's site traffic to be redirected for this experiment is purely driven by the risk tolerance of the experimenter. I feel that a 100% share towards this experiment gives a good and tolerable balance between the length of the experiment of 18 days [17.1 days rounded up] and the risk tolerance of exposing users to uncertain changes. That is, with all traffic redirected as test subjects, there will be 50-50 split between the control and experimental groups where each have 50% of the overall site traffic. If this experiment turns out to have a negative impact on the business, only 50% of all site visitors are at risk. Reducing this risk will require lengthening the duration of the experiment from 18 days, which is not desirable for an A/B test

Experiment Analysis

Sanity Checks

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.

The following values are from sanity sanity check

Number of cookies: [.4988, .5012]; observed .5006; PASS

Number of clicks on “Start free trial”: [.4959, .5041]; observed .5005; PASS

Click-through-probability on “Start free trial”: [.0812, .0830]; observed .0822; PASS

Result Analysis:

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.

Gross Conversion:[-.0291,-.0120],Statistically and Practically Significant

Net conversion: [-.0116, .0019], not statistically significant, not practically significant

Sign Tests

I used the online calculator to perform sign test.

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.

Gross Conversion:0.0026,statistically significant

Net conversion: .6776, not statistically significant

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.

The Bonferroni correction was not used in the analysis phase because our launch decision is based upon the significance of two metrics rather than just one. Had we used just one metric out of several to base our launch decision, the Bonferroni method would be appropriate. But, because the nature of our hypothesis requires that two effects be considered, we cannot base our decision in one metric alone. The sign tests allow for an additional form of analysis. The conclusion from the sign test mirrors that of the effect size test, that gross conversion is significant but net conversion is not. Had we any discrepancies with regard to the significance of the evaluation metrics between the sign and effect size tests, further study would be warranted. In this case, both tests agree and our conclusions with regard to both metrics are strongly supported.

There was no discrepancy between the hypothesis tests and the sign tests.

Recommendation

Make a recommendation and briefly describe your reasoning.

Based on the analysis above

Gross Conversion: The Gross Conversion definition is it is the ratio of the number of users enrolling in the course to the number of user who clicked Start Now Button. As the pop-up page recommend the minimum time required per week to complete this course, it reduce the total number of users enrolling for free Trial. Also the coaches able to concentrate on less number of students, and able to convert them from Free Trial to Paid Service.

Net Conversion: The Net Conversion definitionis the number of users who enrolled for the Free Trial and make their first payment to the number of users who clicked the start free trial button. There is no statistically change in it

My recommendations on this is that the Gross Conversion actually showing positive results and frustrated students left because they don't have enough time. In Net Conversion there is no statistically significant change, but confidence interval does include the negative of the practical significance boundary. There is possibility that the number went down to an amount that matter to business. So, launching the project not going to be helpful.

Follow-Up Experiment:

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.

This Experiment is based on student that is about to pay their first payment, who are in free service.

For Reducing the attrition rate in early of course, that is those students who left the course early because they are unknowledgeable about pre requisite leanings.

For that we need a course or a Project during the Free Trial, so the users able to get an idea what prerequisite knowledge they should know and what the course is all about. The system also need an auto grader so Coaches spend more time to help those users who passed this project and continued in the Course or are in Paid Service.

The motive of this Hypothesis is that introducing the prerequisite course Udacity will be able to know the creamy layer of users and Coaches able to devote more time on those users who are actually interested to pass the Course and get the certificate. The unit of diversion would be user-id as they are already registered in Free Trial.

Invariant Metrics:

Number of users enroll for Free Trial: As the purposes and target of this experiment on those who already enrolled in free Trial.

Evaluation Metrics:

Net Conversion : The net conversion maybe drop or remain low because of PreCourse Project, as this passes by those who clear it, and only allow this creamy layer to go further and explore.

Resources

- http://graphpad.com/quickcalcs/binomial1.cfm
- http://getdatadriven.com/ab-significance-test
- http://www.evanmiller.org/ab-testing/sample-size.html#!16.5;80;5;1;1