A/B Testing Project

夏强
huakaijianwo2009@163.com

Experiment Design

Metric Choice

- Invariant Metrics :

Number of cookies,
Number of clicks
these 2 metrics should not be changed by the experiment ,Number of cookies is used as unit of diversion.

- Evaluation Metrics: Gross conversion, Net conversion

Gross conversion metric implies the experiment's influence on the enrollment of people who click the button.
Net conversion implies the experiment's influence on the payment of people click the "Start free trial" button
they are primary object of this experiment

- others

Click-through-probability:
it is does not depend on our test which is an excellent invariant metric.
but the number of cookies & number of clicks are already sufficient to use as invariant metric, extra metric is not needed.
Number of user ids:
There may be some time that people chose not to enroll have no Udacity count, they likely not to register. And people who chose yes must have a count. So it may be influenced by the experiment. The number of user IDs is usable as evaluation metric because it reflects whether we will reduce the number of students to continue past the free trial. It is not the best metric as it is not normalized (so Gross Conversion is definitively better) but it could potentially be an evaluation metric.
Retention:
Retention may be influenced by the experiment; the experiment may make some people do not complete check out.
Retention has use user ids count as denominator, so it is not a good evaluation metric for this test, though it may be an interest metric to look at in other settings.

- results we look for

Gross conversion:
To test the first part of our hypothesis (reducing students), we want to see whether the experiment group have practically significant lower Gross conversion than the control group, since the students enrolled in free trial aware the time commitment upfront.
Net conversion:
To test the second part of our hypothesis (without significantly reducing the number of students to continue past the free trial), we want to see whether the experiment group do not have significant lower Net conversion than the control group.

Measuring Standard Deviation

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

SE =((p*(1 − p)/n)**0.5

for 5000 page views

Gross conversion

n = 0.08* 5000 = 400

p = 0.20625

SE = 0.0202306

Net conversion

n = 400

p = 0.1093125

SE = 0.01560154

Since the denominator for both two metrics is the number of unique cookies to clicks the button, which is unit of diversion, thus we could expect that the analytical variability would match with the empirical variance.

Sizing

Number of Samples vs. Power

use http://www.evanmiller.org/ab-testing/sample-size.html

$1-\beta = 80\%$

$\alpha = 5\%$

For Gross conversion

dmin= 0.01

p = 0.20625

needed click 25,835

needed pageviews (25,835/0.08)*2 =645876

For Net conversion

dmin= 0.0075

p = 0.1093

needed click 27,411

needed pageviews 685275

Duration vs. Exposure

risk assessment: Indicating needed time commitment is pretty harmless, it can stop student from, and the data is not sensitive neither.

As the experiment is not risky and we should divert the entire traffic 40000

685275/40000 = 17.13188

so, we need 18 days

Experiment Analysis

Sanity Checks

- Number of cookies:

experiment group:344660 control group:345543.
SD =(0.25/(344660+345543))**0.5 = 0.0006018407
0.95 confidence interval
[0.5 - SD*1.96 , 0.5+SD*1.96] = **[0.4988 ,0.5012]**
observed
345543/(344660+345543) =**0.5006**; **Pass**

- Number of clicks

experiment group:28325 control group:28378.
SD =(0.25/(28325+28378))**0.5 =0.002099747
0.95 confidence interval
[0.5 - SD*1.96 , 0.5+SD*1.96] = **[0.4959 ,0.5041]**
observed
28378/(28325+28378)= **0.5005; Pass**

Result Analysis

Effect Size Tests

- Gross conversion:

experiment group 3423/17260
control group 3785/17293
SE = 0.004371675385
m = SE * 1.96 = 0.00856848375
Pooled Probability = 0.2086
D hat = -0.02055
Confidence Interval(CI) = [-0.0291, -0.0120]
Statistically and Practically significant (CI doesn't contain zero, d_min)

- Net conversion:

experiment group 1945/17260
control group 2033/17293
SE = 0.003434133513
m = SE * 1.96 = 0.0067
Pooled Probability = 0.2086
D hat = -0.0049
Confidence Interval = [-0.0116, -0.0018]
Statistically and practically insignificant

Sign Tests

- Gross conversion metric:

There are 4 out of 23 days on which the gross conversion is higher in the experiment group, the two-tailed p value for this sign test is 0.0026
statistically and practical significant

- Net conversion metric:

There are 10 out of 23 days on which the net conversion is higher in the experiment group, using the online calculator, the two-tailed p value for this sign test is 0.6776
statistically and practically insignificant

Summary

Chi-square hypothesis tests and the sign tests results match

we did not use Bonferroni correction, because we need *all* metric match our expectations, Bonferroni correction can be used if we need *any* metric match our expectations.

Recommendation

Udacity should not use the change.
The result shown that Gross Conversion was practically significantly decreased;
This is a good news for Udacity team since coach can now focusing on more quality students.
As for the Net Conversion there has been no statistically significant change, but the confidence interval does include the negative of the practical significance boundary. That is, it's possible that this number went down by an amount that would matter to the business. It would be risky if we launch the change. So, Udacity should not use the change.

Follow-Up Experiment

Experiment: Individualized courses table
Student from different background need different strive to completed the project, before a student start a project, they are offered an exam which is designed to detect his weak point for this project, then based on this exam, the class material was classified to needed course, recommended courses and optional courses.

Null hypothesis: Use the individualized courses table, it will not increase Retention by a practically significant amount.

The new course table will be randomly assigned to a Control & Experiment Group

unit of diversion: Number of user-id
as this experiment only influence users who have sign up for free trial already.
invariant metric: Number of user-ids
because the users sign up for the free trial before get the new course table.
evaluation metric: Retention
which, if positive and practically significant, will show an increase in revenue resulting from this change.
If Retention is positive and practically significant at the end of the experiment, we can launch the new feature.