A/B Testing

Wang Lu

Overview

An online education website Udacity is considering to add a Free Trial Screener where the student clicked "start free trial". Those who indicated fewer commitment than expected for successful completion would be suggested to access free materials instead of continuing free trial enrollment.

Hypothesis was that expectations made clear for students upfront will reduce those who left the free trial simply because of poor time arrangement, without impact on the number of students continue past the free trial and complete the course.

The promising benefits are that Udacity could improve the overall student experience and improve coaches' capacity to support students who are likely to complete the course.

The unit diversion is a cookie. But from the point when students enroll in free trial, they are tracked by user-id. Notice that the uniqueness of cookies is determined by day. In other words the same cookie visiting on different days would be counted twice.

Experiment Design

Metric Choice

I would use **number of cookies, number of clicks, click-through-probability** as invariant metrics; and use **conversion and net conversion** as evalution metrics. Retention is not directly relevant to our hypothesis so I would rather not use it.

To indicate a clear structure of events, I divided the conversion funnel into 5 phases A, B,C,D,E, which assumed to happen sequentially: A. View the course overview --> B. Click 'start free trial' button --> C. Finish checkout and start free trial (**Free Trial Screener might be added between**) --> D. Remain enrolled and make at least one payment --> E. Complete the course. Number of Cookies. Number of Clicks, Number of User-ids are metrics for user involement in phase A, B, C respectively. While Click-Through Probability, Gross Conversion, Retention, Net Conversion stands for conversion rate from phase A to B, B to C, C to D an D to E respectively.

- ✓ **Number of Cookies** (the number of cookies to view the course overview): It definitely should be an **invariant metric** which supposed to be independent from the change we make. Because it happens before users click the 'start free trial button', the event after which we introduce the change. Besides, cookie as the unit of diversion we use initially, is important reference we'd like to look at sanity check.
- ✓ **Number of clicks** (the number of unique cookies to click 'start free trial' button at least once): Again, a very good **invariant metric** since it is independent of change. We want to make sure a consistent Number of Clicks between experiment group and control group.
- ✓ **Click-through-probability** (the number of unique cookies to click the start free trial button divided by number of unique cookies to view the course overview page): It is invariant between 2 groups thus could be treated as an **invariant metric.** Besides, as it normalizes to the size of the control and experiment group, it is a even better invariant metric compared to the number of clicks.
- ✓ **Gross conversion** (the number of user-ids enrolled in free trial divided by number of unique cookies to click the 'start free trial'): Similarly it should be treated as **evaluation metric** to test the part of hypothesis that Udacity would improve student experience and less frastrated student leave after free trial.
- ✓ **Net conversion** (the number of user-ids remain enrolled after free-trial divided by the number of unique cookies to click the 'start free trial' button): It is good **evaluation metric** in that it is direct indicator of whether launch of Free Trial Screener would reduce website revenues. Even though Udacity cares about improving the experience of students who are more likely to complete the course, it should be in the condition that net conversion shouldn't decrease significantly .

Why not choose the others as metrics?

- Number of user-ids (the number of users who enroll in the free trial): If hypothesis held true it should reduce, thus it can't be used as invariant metric. Surely we can use it as evaluation metric because it's sensitive to change. However we already have net conversion to inform us about the second part of our hypothesis, which is better since it normalizes between groups, we can simply drop this rebundant one.
- Retention (the number of user-ids remain enrolled after free-trial and thus make at least one payment divided by number of userids to complete checkout): as a downstream metric of numbers of user-id, it's definitely not suitable as invariant metric. And as it is not directly related to hypothesis, making it a bad candidate for evaluation metric as well.

When to Launch?

If the implement does impact user behavior as expected, I would expect a significant drop in Number of User-ids since users without enough commitment would be discouraged from enrollment thus less user-ids created. And if the improvement can be achieved without the expenses of website revenue, the effect of reduced Number of User-ids and user-ids remain enrolled after free-trial would be expected cancell each other out and as results it produces a non-significantly decreased Net Conversion. Therefore to launch the change, two conditions should be met: **a practically significant drop in Gross Conversion** (less unqualified students enrolled) and **none practical significant change in Net Conversion** (no less completion).

Measuring Standard Deviation

Given that 5000 of unique cookies view course pages, and assuming binominal distribution for Gross Conversion, Net Conversion and number of user-ids divided by unique cookies viewing course pages,

- **For Gross Conversion:**
  N=3200/8=400
  Baseline value (given)=0.20625
  SE= sqrt(0.20625*(1-0.20625)/400)= 0.0202
- **For Net Conversion:**
  N=3200/8=400
  Baseline value (given)= 0.109313
  SE=sqrt(0.109313*(1-0.109313)/400)= 0.0156
- **For Number of User-ids:**
  N=5000
  Baseline value =660/40000=0.0165
  SE=sqrt(0.0165*(1-0.0165)/5000)=0.0018

The analytical estimates of **all of the 3 metrics suppose to be comparable** to the empirical variability, considering they all use unit of diversion as denominator (unique cookies)

Sizing

Number of Samples vs. Power

Using the web based calculator, http://www.evanmiller.org/ab-testing/sample-size.html, I got the required size for the 3 of evaluation metrics respectively.

- **For Gross Conversion:**
  Beta(given)=0.2
  Alpha(given)=0.05
  Baseline value (given)=0.20625
  dmin(given)=0.01
  sample_size (unique cookies click 'Start Free Trial' button)=25835

Considering click-through probability(CTP=0.08) and setting of control group, the number of cookies viewing the course page should be at least: 25835/0.08*2=645875. Given average course page viewing traffic is 40000, 17 days is required

- **For Net Conversion:**
  Beta(given)=0.2
  Alpha(given)=0.05

Baseline value (given)= 0.109313
Dmin=0.0075
Sample_size(unique cookies click ‘Start Free Trial’ button)=27413

Similarly, Number of pageviews=27413/0.08*2=685325
685325/40000=17.133125, at least 18 days is required.

- **For Number of User:**
  Beta(given)=0.2
  Alpha(given)=0.05
  Baseline value(calculated) =0.0165
  Sample_size(unique cookies click ‘Start Free Trial’ button)=164843

Number of pageviews=164843/0.08*2=4121075
4121075/40000=103.03, thus 104 days is required. **It’s not realistic to conduct such long experiment. Thus I decide to drop this metric and go ahead with only Gross Conversion and Net Conversion**. As we can see from definition the two are highly positively correlated. Therefore Bonferroni correction shouldn’t be conducted otherwise the results would be too conservative.

Duration vs. Exposure

Considering it is just add a pop-up survey windows there is low technical risk. Besides no sensitive information is collected and the data is adequately aggregated and not identifiable. I would rather divert all traffic into experiment considering without asking for consent. If so I would need 18 days to do the experiment which is feasible devotion.

Experiment Analysis

Sanity Checks

- For Number of Cookies:
  P(probability of a unique cookie in control group)=0.5
  SE = sqrt(0.5*(1-0.5)/(345543+344660))= 0.0006018
  M(margin of error)= 0.0006018*1.96=0.0012

CI(confidence interval)=[0.5-0.0012,0.5+0.0012]= [0.4988,0.5012]
Observed Value=345543/(345543+344660)=0.5006
Since observed value fall within CI, it passes sanity check.

- For Number of Clicks:
  P(probability of a unique cookie in control group)=0.5
  SE = sqrt(0.5*(1-0.5)/(28378+28325))= 0.0020997
  M(margin of error)= 0.0020997*1.96=0.0041

CI(confidence interval)=[0.5-0.0041,0.5+0.0041]= [0.4959,0.5041]
Observed Value= 28378/(28378+28325)=0.5005
Since observed value fall within CI, it passes sanity check.

Result Analysis

Effect Size Tests

- **For Gross Conversion, from day1 to day23:**
  Control group: Sum(unique cookies click ‘Start Free Trial’)=17293
  Sum(users get enrolled in free trial)=3785
  Pexp=3785/17293=0.2189
  Experiment group: Sum(unique cookies click ‘Start Free Trial’)=17260
  Sum(users get enrolled in free trial)=3423
  Pcon=3423/17260=0.1983

d =Pexp-Pcon=0.1983-0.2189= -0.0206
Ppool=(3785+3423)/(17293+17260) = 0.2086
SE = sqrt(0.2086*(1-0.2086)*(1/17293+1/17260))= 0.0043716
M=0.0043716*1.96=0.0086
Confidence Interval = [-0.0206-0.0086, -0.0206+0.0086]=[-0.0291,-0.0120]

Since both 0 and -d_min(-0.01) fall outside the upper bound of CI, we can conclude that Gross Conversion in experiment group is both statistically and practically significant.

- **For Net Conversion, from day1 to day23:**

  Control group: Sum(unique cookies click ‘Start Free Trial’)=17293
  Sum(users remained after free trial)=2033
  Pexp=2033/17293=0.1176

  Experiment group: Sum(unique cookies click ‘Start Free Trial’)=17260
  Sum(users remained after free trial)=1945
  Pcon=1945/17260=0.1127

  d =Pexp-Pcon=0.1127-0.1176= -0.0049
  Ppool=(2033+1945)/(17293+17260) = 0.1151
  SE = sqrt(0.1151*(1-0.1151)*(1/17293+1/17260))= 0.0034338
  M=0.0034338*1.96=0.0067
  Confidence Interval = [-0.0049-0.0067, -0.0049+0.0067]= [-0.0116,0.0018]

Since both 0 and -d_min(-0.0075) are included by CI, thus it is neither statistically significant nor practically significant. However, we should note that there is a risk for business having the net conversion dropped (i.e., losing revenues) if launch the change, since confidence interval includes the lower practical significant boundary.

Sign Tests

Assuming a binominal distribution for signs with baseline value equals to 0.5:

- **For Gross Conversion, from day1 to day23:**
  Only 4 days the Gross Conversion in experiment group is greater than that in control group. Assuming a binominal distribution of signs with a baseline value=0.5, P(4 positives|23 trials)=0.0026, lower than alpha, null hypothesis is rejected. Gross Conversion in experiment group is significantly smaller than that in control.

- **For Net Conversion, from day1 to day23:**
  For 10 out of 23 days the Net Conversion in experiment group is greater than that in control group. Assuming a binominal distribution of signs with a baseline value=0.5, P(10 positives|23 trials)=0.6776 >>alpha, fail to reject null hypothesis. Net conversion in experiment group is not significantly different from that in control.

Both of the sign tests agree with respective effect size tests.

Summary

I chose not to use Bonferroni correction due to 2 reasons. First the 2 metrics we use are highly positively correlated. The greatest criticism for using Bonferroni correction is its over-conservativeness under this situation. (https://en.wikipedia.org/wiki/Bonferroni_correction)

Recommendation

I recommend **not to launch the change right now**. Since according to our hypothesis, it is beneficial to launch the change when the two conditions, statistically and practically significant decrease of gross conversion and none significant decrease of net conversion, are both met. Now we find it is true based on our analysis above. It seems that we should launch it without hesitation.

However, we should notice that there is a risk for business losing revenues if it launches the change, since confidence interval of net conversion includes the lower practical significant boundary. It means we can't confidently prove launch of screener can improve user experience free of the risk of losing potential customers. Therefore, I recommend not to launch the change right now. Except Udacity collects more information for determining whether the negative impact indeed exists and how great it is. Till then the personnel from user experience section and conversion section can sit together and redecide whether or not to launch the Free Trial Screener.

Follow-Up Experiment

Instead of doing a subtraction by keeping s out unqualified students, for me it's a better thinking that how we can hold more students after free trial. It's very true that students get frustrated and quit if they have no idea when can they finish the course. So we might introduce at the beginning of free trial a customized learning schedule generated based on each one's background, allocable time as well as their later performance in free trial, which inspires them to complete learning by leading them foreseeing their future achievement and when they will make it.

In sum, my hypothesis is that customized learning schedule across free trial inspires students to keep going and complete courses, therefore decrease dropping out students after free trial. (The situation of whether student actually finish courses is too complex to test here as students might stop and resume subscription based on their schedule and we can't conduct A/B test over a unsure or too long time span.)

My **unit of diversion** is user-ids, as we are interested only in user behavior who enrolled by free trial.

The invariant metrics is user-ids as it checks whether the control and the experiment groups are powered by the same number of subjects across experiments, therefor make sure results from two groups are comparable.

The evaluation metric is Retention (the number of user-ids remain enrolled after free-trial and thus make at least one payment divided by number of userids to complete checkout).

As long as statistically and practically significant increase is found in retention is found, Udacity should consider implement this change.

Resources

Udacity Forum Discussion

Online lecture notes: http://napitupulu-jon.appspot.com/categories/ab-testing.html

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

https://en.wikipedia.org/wiki/Bonferroni_correction