Patrick Coakley
Udacity Data Analytics - P7
6/18/2017

EXPERIMENT DESIGN

Metric Choice

The goal of this experiment is to reduced the number of frustrated students by implementing a screener that will recommend that at least 5 hours a week is needed to be a successful nanodegree student at Udacity. We also want to do this without reducing the number of students that stay on after the free trial, thus becoming paid customers. The unit of diversion for this experiment is cookies. The evaluation metrics that correlate closest to the goals of the experiment are Gross Conversion and Net Conversion.

Gross Conversion, which is the number of users that complete the checkout process to enroll divided by the number of unique cookies to click the free trial link, measures the number of students that signed up for the free trial from the link, and thus is directly effected by the screener; it correlates with the goal of reducing the number of students that become frustrated with the free trial. The optimal outcome of this experiment would be for Gross Conversion to be reduced.

Net Conversion, which is the number of users that complete the checkout process to enroll divided by the number of unique cookies to click the free trial link, correlates to the secondary goal of not reducing the number of enrolled students that stay on for payment. Thus, we want Net Conversion to go up or, at the very least, stay the same.

Invariant Metrics:

- **Number of cookies** - the number of cookies are not going to be effected by the free trial experiment because the cookies are stored when the site loads, so it would have no impact on the results as anyone, paying or not, would store the cookie

- **Number of clicks** - similar to the number of cookies, the number of clicks does not change during the experiment because users will already be clicking the free trial link if they have interest in it, so any changes in the experiment would only be measured after clicking it

- **Click-through profitability** - taking the number of unique cookies that click the free trial link and dividing it by the unique cookies that view the course overview page is not going to change throughout the testing because the user would be clicking the free trial link regardless of the screener

Evaluation Metrics:

- **Gross conversion** - the number of users that complete the checkout process to enroll divided by the number of unique cookies to click the free trial link is a useful evaluation metric because you can see how many users started the free trial after implementing the new screener

- **Net conversion** - taking the retained users (those that remained enrolled and thus paid) and dividing them by the number of unique cookies that clicked the free trial link gives a nice ratio of how many paid users you have against the total number of users that clicked the trial but for one reason or another either did not stay past the free trial or never completed checkout in the first place

Unused Metrics:

- **Number of user-ids** - the total number of users that enroll in the free trial is not very useful for either type of metric, as it would actively change between the two groups, making it too variable to be invariant, and, in addition, gross conversion already covers enrollments and thus it would be redundant in this experiment

- **Retention** - the number of users that remain enrolled after the free trial divided by the total number of users is a good indicator of how much the changes would effect the number of new users, positively or negatively, but for this experiment it is not a good fit due to the time-sensitive nature of the launch criteria, as seen in the Sizing section below

Measuring Standard Deviation

Using the numbers from the baseline values document we can apply the Bernoulli distribution formula of sqrt(p * (1 - p) / N) to solve for the standard deviations.

Gross Conversion

With the Gross Conversion, p is the probability of enrolling after clicking the free trial button, while N can be defined as the click-through probability rate of the free trial link times the sample size:

p = 0.20625
N = 0.08 * 5000 = 400
sqrt((0.20625 * (1 - 0.20625)) / 400) = **0.020230604**

Retention

With the Retention, p is the probability of remaining enrolled and thus paying, while N can be defined as the enrollment rate of the free trial link times the sample size:

p = 0.53
N = 0.0165 * 5000 = 82.5
sqrt((0.53 * (1 - 0.53)) / 82.5) = **0.054949012**

Net Conversion

With Net Conversion, p is the probability of remaining enrolled past the free trial, while N can be defined as the click-through probability rate of the free trial link times the sample size

p = 0.1093125
N = 0.08 * 5000 = 400
sqrt((0.1093125 * (1 - 0.1093125)) / 400) = **0.015601545**

Since the unit of diversion is cookies and the unit of analysis was cookies, it makes sense that they would be comparable to the empirical results.

Sizing

Number of Samples vs. Power

The following were calculated using Evan's Awesome A/B Tools. The Bonferroni correction will not be used:

Gross Conversion

- dmin = 0.01
- rate = 0.20625
- alpha = 0.05
- beta = 0.2
- pageviews = 25835
- clickthrough rate = 0.08
- total pageviews = (2 * 25835) / 0.08 = **645875**

Retention

- dmin = 0.01
- rate = 0.53
- alpha = 0.05
- beta = 0.2
- pageviews = 39,115
- clickthrough rate = 0.0165
- total pageviews = (2 * 39115) / 0.0165 = 4741212.121212121 => **4741212**

Net Conversion

- dmin = 0.0075
- rate = 0.1093125
- alpha = 0.05
- beta = 0.2
- pageviews = 27413
- clickthrough rate = 0.08
- total pageviews = (2 * 27413) / 0.08 = **685325**

If Retention was included, it would require **4741212** page views. However, it will not be used as seen in the next section.

As Net Conversion required the larger number of page views, the total for both is **685325**.

Duration

To calculate the duration of the experiment, we take the number of page views from the last section and divide it by the number of unique cookies to view a page per day.

If we were to include Retention as a metric, it would take: 4741212 / 40000 = 118.5303 => **119 days**

This is outside of the scope of the experiment, as it was mentioned that it should only take "a few weeks" as stated in the original document. Therefore, Retention is not going to be used.

Since Net Conversion required the most days of the two remaining metrics, it would require:

685325 / 40000 = 17.133125 => **18 days**

As this is a very short duration, it makes sense to use a fraction of 1 to divert because it will allow for faster results in a very short experimentation period. There are no risks involved in doing the experiment in this way, as the experiment itself is simply a measure of whether or not users opted sign up for the free trial, and potentially stay on as a paying customer. No sensitive information would be collected during the experiment, and outside of reporting the aforementioned statistics Udacity would run ostensibly the same.

EXPERIMENT ANALYSIS

Sanity Checks

Using the final project results document, we can use the values to calculate sanity checks on the invariant metrics. Since there is a control group and an experimental group, the probability is 0.5, and therefore the z-score is 1.96 (except in the case of click-through profitability).

Number of Cookies

- stdev = sqrt((0.5 * 0.5) / (345543 + 344660)) = **0.000601841**
- me = 0.000601841 * 1.96 = **0.001179608**
- lowerbound = 0.5 - 0.001179608 = **0.498820392**
- upperbound = 0.5 + 0.001179608 = **0.501179608**
- observed = 345543 / (345543 + 344660) = **0.500639667**
- passed = **true**

Number of Clicks

- stdev = sqrt((0.5 * 0.5) / (28378 + 28325)) = **0.002099747**
- me = 0.002099747 * 1.96 = **0.004115504**
- lowerbound = 0.5 - 0.004115504 = **0.495884496**
- upperbound = 0.5 + 0.004115504 = **0.504115504**
- observed = 345543 / (345543 + 344660) = **0.500467347**
- passed = **true**

Click-through Profitability

- stdev = sqrt(0.082154091 * (1 - 0.082154091) * (1 / 345543 + 1 / 344660)) = **0.000661061**
- me = 0.000661061 * 1.96 = **0.001295679**
- d = 0.082154091 - 0.082182441 = **-2.83498E-05**
- lowerbound = -2.83498E-05 - 0.001295679 = -**0.001324029**
- upperbound = -2.83498E-05 + 0.001295679 = **0.001267329**
- observed = 0.082125814 - 0.082182441 = **0.000056627**
- passed = **true**

Effect Size Tests

Using the same document as before, we can calculate the effect size tests as well. The Bonferroni technique was not used in this experiment. It is used to "adjust probability (p) values because of the increased risk of a type I error when making multiple statistical tests." We can define a type I error as a false positive, or rather, "[w]hen the null hypothesis is true and you reject it."

Using the Bonferroni Technique can increase type II errors, which is a situation in which "the null hypothesis is false and you fail to reject it." Our hypothesis is hinged on two outcomes: we must have the number of frustrated students decrease while keeping the number of enrolled users the same or higher. Therefore it makes more sense to forgo the Bonferroni Technique in this situation, as it would possibly make it too difficult for both parameters to be successful; we need both of these to come true in order for our experiment to succeed.

Gross Conversion

- dmin = 0.01
- ppool = (3785 + 3423) / (17293 + 17260) = **0.208607067**
- sepool = sqrt(0.208607067 * (1 - 0.208607067) * (1/17293 + 1/17260)) = **0.004371675**
- d = (3423 / 17260) - (3785 / 17293) = **-0.020554875**
- m = 1.96 * 0.004371675 = **0.008568484**
- lowerbound = -0.020554875 - 0.008568484 = **-0.029123358**
- upperbound = -0.020554875 + 0.008568484 = **-0.011986391**

These results are both **statistically significant** and **practically significant** because they do not contain zero, nor do they contain the dmin.

Net Conversion

- dmin = 0.0075
- ppool = (2033 + 1945) / (17293 + 17260) = **0.115127485**
- sepool = sqrt(0.115127485 * (1 - 0.115127485) * (1/17293 + 1/17260)) = **0.003434134**
- d = (1945 / 17260) - (2033 / 17293) = **-0.004873723**
- m = 1.96 * 0.003434134 = **0.006730902**
- lowerbound = -0.004873723 - 0.006730902 = **-0.011604624**

- upperbound = -0.004873723 + 0.006730902 = **0.001857179**

These results are **not statistically significant** because they contain zero, and therefore would be deemed **not practically significant**.

Sign Tests

For the calculation of the sign tests, I used this website. The probability was 0.5 as there are two groups, and there were 23 recorded payments and enrollments, so only the clicks up to their respective measurements were used.

Gross Conversion

Four out of twenty-three results for the experiment were successful, giving a twin-tailed P of **0.0026**.

This is **statistically significant** because it is lower than the alpha, 0.5.

Net Conversion

Ten out of twenty-three results for the experiment were successful, giving a twin-tailed P of **0.6776**.

This is **not statistically significant** because it is greater than the alpha, 0.5.

Recommendation

Based on the criteria for launching the experiment, as well as the measures of success set in the hypothesis, it would not make sense to launch the experiment at this time. The original hypothesis states that the experiment "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." Ultimately the goal is to provide a better student experience, but without hurting business.

This correlates to our two evaluation metrics, with gross conversion being relative to the number of enrollments per click, and net conversion being relative to payments after the free trial per click. From the results we can see that the gross conversion had the desired effect of reducing the number of free trials, but the net conversion automatically fails to meet the criteria set out from the beginning because it falls below zero, and thus reduces the number of students continuing past the free trial; the partially positive results are too weak to consider. Thus, it would make the most sense to take a step back and re-evaluate the experiment.

Factors that might come into question include the duration of the free trial, the number of recommended hours, and method of presentation for the screener itself. I think incorporating testimonials of busy people with little time also might give a better barometer of how much time is necessary versus the single-sentence description given in the example. It would be very easy to re-launch using the same set of evaluation metrics while tweaking the above changes to the experiment.

Follow-up Experiment

For a follow-up experiment to help reduce the number of early cancellations during the trial period, I would propose offering the student an optional one-on-one student mentor appointment from the start, prior to starting any classes, and making sure to highlight it as a part of the program. With Massive Open Online Courses, or MOOCs, I think there is a tendency to feel alone, and I can understand a lot of frustration with students who are in a free trial for a paid program and having a feeling of anxiousness about learning something new. Udacity is unique in that many other programs like it don't offer this level of support, and I think it makes sense to incorporate it from the very start. Making the student mentor feature a part of the trial helps fix a few issues:

1. It lets students know the feature exists, making them more likely to use it from that point forward.
2. It helps students that feel like they're alone have some support at the beginning.
3. It gives students a chance to discuss what they're learning in a context outside of just the exercises.

From this we can infer that hypothesis of the experiment would be:

H0 - The mentor appointments do not reduce the number of cancellations during the trial period.

Ha - The mentor appointments reduce the number of cancellations during the trial period.

Since we are measuring enrolled students, the unit of diversion should be users, with the invariant metric being the total number of user-ids. This is because we are measuring students who have already enrolled, and everything past that would not be related to clicks or cookies, and only user-ids. For evaluation, it would make sense to use the retention rate, as this measures the number of students who remained enrolled. Using these we could get an idea as to whether or not the number of paid students is growing based on the incorporated changes.

Sources

- **Siroker, Dan, and Pete Koomen**. *A/B Testing: The Most Powerful Way to Turn Clicks into Customers*. Hoboken, NJ: Wiley, 2013.
- Udacity Forums
- When to use the Bonferroni correction
- What are type I and type II errors?