A-B-Test---Kristy Morris Project 7--Udacity

Metric Choice--Gross Conversion/Retention/Net Conversion

The metrics I chose are listed below and why--

Gross conversion: That is, The numbers of user-ids who decide to start the free trial and are expected to depend on how the "start free trial" page is carried out --- whether a "5 or more hours per week" is suggested --- this is one question we would like to understand through this A/B test. Therefore, this is a good evaluation metric.

Retention: Along the same lines, as the above, we would like to understand whether carrying out a "5 or more hours per week" suggestion is helpful to increase ratio of user-ids who will make payments over those who finish the free trial, and therefore this metric is good for evaluation.

Net conversion: That is, The result of the previous two evaluation metrics: gross conversion and retention, and it can be considered as a more general goal of the A/B test --- whether carrying out 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, again a good evaluation metric.

***Invariant Metrics--

Number of cookies--we do not want the number of users that visit the website to differ as we carry out the "start free trial" page. The users have not seen this page before. Independent of the experiment/test--Resulting in this one being a invariant metric.

Number of user-ids--this one is neither invariant nor evaluation metric. Due to the enrollment could be dependent upon how we carry out the "start free trial" page, we could expect to see different values in the control and experiment group. Also, the number of visitors between experiment are likely to be different & number of enrolled user can fluctuate in a particular day, which likely could skew the results.

Number of clicks-- Comparable to number of cookies, this metric does not depend on how we carry out the "start free trial" page, since the clicked users have not seen that

page before they decide to click the button. Example-- the page asking the number of hours that the student can dedicate to course work after clicking "Start free Trial" button, but the course overview page remains the same for both the control & the experiment/test group. Resulting in this one being a invariant metric.

Click-through-probability--Again, since the users have not seen the page we tested on before they decide to click the button, the click-through-probability also does not depend on our test. This one would be an invariant metric. This is a real good invariant metric, since the clicks happen before the users see the experiment, therefore it does not depend on our test.

***Launch Criteria Statement-- The expectation for the experiment/test could be as follows--the gross conversion will decrease practically significance, which demonstrates that the cost will be lower by introducing the new screener; while the net conversion will not decrease statistically significance, which can indicate that the screener can impact the result of the revenues. Bonferroni correction could be use as a method of tracking multiple metrics--to assist with ruling out false positives, however it is a conservative method.

Measuring Variability--

The analytical standard deviation is computed as std = sqrt(p * (1-p)/N)

Gross conversion: std = sqrt(0.20625*(1-0.20625)/3200) = 0.00715 (correspond to 3200 clicks & 40000 pageviews). For 5000 pageviews, we have new std = 0.00715 * sqrt(40000/5000) = 0.0202

Retention: std = sqrt((0.53*(1-0.53)/660) * sqrt(40000/5000))) = 0.549

Net conversion: std = sqrt(0.1093125*(1-0.1093125)/3200) = 0.0055159 (correspond to 3200 clicks & 40000 pageviews). For 5000 pageviews, we have new std = 0.0055159 * sqrt(40000/5000) = 0.0156

Gross conversion: 0.0202 Retention: 0.0549 Net conversion: 0.0156

For each metric you selected as a evaluation metrics, indicate whether you think the analytic estimate would be comparable to the the empirical estimate of the variability, if you had time? Expained below--

For both Gross Conversion and Net Conversion using number of cookies as denominator, which is also unit of diversion. Here, the unit of diversion is equal to unit

of analysis, which indicate the analytical estimate would be comparable to the emperical variability.

For Retention, the denominator is "Number of users enrolled (complete checkout)" which is not similar as Unit of Diversion. The unit of analysis and the unit of diversion are not the same therefore the analytical an the empirical estimates are different.

Sizing--

Number of Samples vs. Power

I decided not to use Bonferroni correction, because the metrics in the test has high correlation and the Bonferroni correction will be too conservative to use it.

I calculated the number of samples needed for each metric using the online calculator, with alpha = 0.05, 1 - beta = 0.2. The baseline conversion rate and minimum detectable effect (d_min) are listed individually below. Also note that the number produced by the online calculator is per branch, and in order to have both control and experiment, we need to double the number of required page views.

Gross conversion = The baseline conversion rate is 0.20625, and d_min is 0.01. The required number of samples calculated from the online calculator is 25,835. Note that this is the number of clicks on "start free trial", and in order to get that number, we need 25,835/ 0.08 * 2 = 645,875 page views.

Retention = The baseline retention rate is 0.53, and d_min is 0.01. The required number of samples calculated from the online calculator is 39,115. Note that this is the number of users who finished the 14 day free trial, and in order to get that number, we need 39,115/ 0.08 / 0.20625 * 2 = 4,741,212 page views.

Net conversion = The baseline conversion rate is 0.1093125, and d_min is 0.0075. The required number of samples calculated from the online calculator is 27,413. Note that this is the number of clicks on "start free trial", and in order to get that number, we need 27,413/0.08 * 2 = 685,325 page views.

If we keep the retention rate as a evaluation metric, the number of required pages will be too large (in order to get 4.7 million page views, it takes 117 days of 100% or full site traffic, which is not realistic). Therefore, we decide to drop the retention rate evaluation metric, and use gross conversion and net conversion as evaluation metrics, and the required number of page views (taking the larger one) is 685,325.

Duration vs. Exposure

I selected to redirect 50% of the traffic to our experiment, and the length of the experiment is therefore 685,325/(40,000 * 0.5) = aprox 35 days (where 40,000 is the baseline number of visitors per day).

The 50% traffic is being redirected to the experiment means that 25% will go to control group and 25% to experiment group, and therefore we risk about a quarter of users seeing an not yet evaluated feature. This could be a financial or business risk, because those 25% users will see a different view of "start free trial" page which potentially discourages them to start the free trial (although the intention is to increase the overall net conversion). However, this is what we will be directly monitoring in this experiment. This choice is needed in order to keep the length of the experiment in a reasonable amount of time. If we reduce the potential risk by half (sending 12.5% users to see not-yet-evaluted feature), the length will be doubled, taking more than 2 months, which is a little too long to run in most cases.

To try and get below 30 days duration if the client suggests--then we could redirect 80% of the traffic to our experiment, and the length of the experiment is therefore 685,325/(40,000 * 0.8) = aprox 22 days. Again, the 80% traffic is being redirected to the experiment and would be = to 40% will go to the control group and 40% to the experiment group. Not a sensitive personal information or medical information issue, so not a risk that we could not monitor/track.

Analysis--

Sanity Checks

For counts ("number of cookies" and "number of clicks"), we model the assignment to control and experiment group as a Bernoulli distribution with probability 0.5. Therefore the standard deviation is std = sqrt(0.5 * 0.5 / (N_1 + N_2)), and the margin of error is me = 1.96 * std. The lower bound will be 0.5 - me and the higher bound will be 0.5 + me. The actual observed value is number of assignments to control group divide by the number of total assignments.

Number of cookies

control group total = 345543 experiment group total = 344660 standard deviation = sqrt(0.5 * 0.5 / (345543 + 344660)) = 0.0006018 margin of error = 1.96 * 0.0006018 = 0.0011796 lower bound = 0.5 - 0.0011797 = 0.4988 upper bound = 0.5 + 0.0011797 = 0.5012 observed = 345543 / (345543 + 344660) = 0.5006 The observed value is within the bounds, and therefore this invariant metric passed the sanity check.

Number of clicks on "start free trial"

control group total = 28378 experiment group total = 28325 standard deviation = sqrt(0.5 * 0.5 / (28378 + 28325)) = 0.0021 margin of error = 1.96 * 0.0021 = 0.0041 lower bound = 0.5 - 0.0041 = 0.4959 upper bound = 0.5 + 0.0041 = 0.5041 observed = 28378 / (28378 + 28325) = 0.5005 The observed value is within the bounds, and therefore this invariant metric passed the sanity check.

Click-through-probability on "start free trial"

For click through probability, we first compute the control value p_cnt, and then estimate the standard deviation using this value with experiment group's sample size, i.e. std = sqrt(p_cnt * (1 - p_cnt) / N_exp). The margin of error is 1.96 times of standard deviation.

control value = 0.0821258 standard deviation = sqrt(0.0821258 * (1-0.0821258) / 344660) = 0.000468 margin of error = 1.96 * 0.000468 = 0.00092 lower bound = 0.0821258 - 0.00092 = 0.0812 upper bound = 0.0821258 + 0.00092 = 0.0830 experiment value = 0.0821824 The observed value (experiment value) is within the bounds, and therefore this invariant metric passed the sanity check.

Effect Size Tests

Please notice that N denotes the number of total samples (denominator) and X denotes the number of target samples (numerator), and _cnt denotes controlled group and _exp the experiment group.

We first computed pooled probability and pooled standard error as

p_pooled = (X_cnt + X_exp) / (N_cnt + N_exp) se_pooled = sqrt(p_pooled * (1-p_pooled) * (1./N_cnt + 1./N_exp)) The probability difference is computed as

d = X_exp / N_exp - X_cnt / N_cnt With these values in hand, the lower bound and upper bound are

lower = d - se_pooled upper = d + se_pooled

Gross conversion

For gross conversion, the total samples (denominator) are the clicks of "start free trial", and the target samples (numerator) are enrolled users.

The caculation is shown below.

N_cnt = clicks_controlled = 17293 X_cnt = enroll_controlled = 3785 N_exp = clicks_experiment = 17260 X_exp = enroll_experiment = 3423

p_pooled = (X_cnt + X_exp) / (N_cnt + N_exp) = 0.2086 se_pooled = sqrt(p_pooled * (1-p_pooled) * (1./N_cnt + 1./N_exp)) = 0.00437

d = X_exp / N_exp - X_cnt / N_cnt = -0.02055

lower = d - se_pooled = -0.0291 upper = d - se_pooled = -0.0120

Since the interval does not contain 0, the metric is statistical significant. It does not include d_min = 0.01 or -d_min = -0.01 either, and therefore it is also practical significant.

Net conversion

For net conversion, the total samples (denominator) are the clicks of "start free trial", and the target samples (numerator) are paid users.

The caculation is shown below.

N_cnt = clicks_controlled = 17293 X_cnt = pay_controlled = 2033 N_exp = enroll_experiment = 17260 X_exp = pay_experiment = 1945

p_pooled = (X_cnt + X_exp) / (N_cnt + N_exp) = 0.1151 se_pooled = sqrt(p_pooled * (1-p_pooled) * (1./N_cnt + 1./N_exp)) = 0.00343

d = X_exp / N_exp - X_cnt / N_cnt = -0.0048

lower = d - se_pooled = -0.0116 upper = d + se_pooled = 0.0019

Since the interval contains 0, it is not statistical significant, and consequently not practical significant either.

Sign Tests

I used the online calculator to perform the sign tests. For gross conversion, the number of days we see an improvement in experiment group is 4, out of total 23 days of experiment. With probability 0.5 (for sign test), the online calculator calculates a p-value 0.0026, which is smaller than alpha = 0.05. Therefore the change is statistical significant.

For net conversion, the number of days we see an improvement in experiment group is 10, out of total 23 days of experiment. With probability 0.5 (for sign test), the online calculator calculates a p-value 0.6776, which is larger than alpha = 0.05. Therefore the change is not statistical significant.

Summary--

I decided not to use Bonferroni correction, because the metrics in the test have a high correlation and the Bonferroni correction will be too conservative to use it. Both the effective size hypothesis tests and sign tests state that the change will practically significantly reduce the gross conversion, however not affect the net conversion rate in a practically significant ways. I completely understand the importance to correct if a test is launched and the metrics show a significant difference, because it's more likely that one of multiple metrics will be falsely positive as the number of metrics increases. However, we would only launch if all evaluation metrics show a significant change. In that case, there would be no need to use Bonferroni correction. We could of also set up our experiment to consider other strategies, such as using a higher confidence level than the 95% used for each metric. Possibly, a 99% confidence level.

Recommendation--

Based on the data analysis above, I recommend not to move forward with the changes of adding "5 or more hour" recommendation to "start free trial" date. The reason is that the A/B test shows that this will not practically significantly increase the net conversion rate. Meaning, it does not increase the number of paid users, which fails the original goal of our experiment/test. In addition, the confidence interval does include a negative number of the practical significance boundary; meaning, it's possible that this number went down by an amount that could hurt the business - decrease the revenue. If we consider the initial hypothesis, it does not increase numbers of paid users, which fails the inital goal of launching this feature and likely to be a unacceptable risk to launch.

Follow-Up Experiment--

I would recommend adding a "accelerated completion program discount option" button on the home page. This would give the users a tuition discount, if they complete the program within a shorten time period, such as 6 months. This feature will be potentially engage the users who are already determined to take the course, and want to jump right in.

The hypothesis is that by providing this additional option, the number of enrollees will increase, because those who decide to take the course will directly enroll rather than experiencing the free trial, during which they might decide not to enroll for certain reasons. Another hypothesis is that this feature will bring more revenue to Udacity --- even though the users who enroll directly pay less than others, the increasing number of users will be more significant. Meaning, larger volume of enrolled users.

Corresponding to each above, there are two evaluation metrics I would use; 1. the conversion rate from home page viewers to the enrolled users. This would test whether the additional option helps to boost the enrollment. 2. the ratio of revenue over number of home page viewers. This will test for the same unit number of users who viewed the home page, whether the additional option helps to increase the overall revenue of Udacity. In addition, 3. retention, the number of user-ids, that clicked the new discount button and stayed enrolled for 14 days and made their first payment, divided by the number of user-ids to complete checkout--this would take some time to measure, but would be good to know.

Invariant metric could be the number of clicks--unquie cookies that clicked on the "accelerated completion program discount option" button that day.

The initial unit of diversion will be a cookie, because the home page viewers are not necessarily signed in. When users are signed in, user id will be used instead of cookies.

References--

Udacity A/B Course Lesson Videos

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

Bernoulli Distribution https://en.wikipedia.org/wiki/Bernoulli_distribution

Feedback from 1st reviewer 3/5/2017