Design an A/B Test

Udacity Data Analyst Nanodegree P7

Experiment Design

Metric Choice: Invariant Metrics

The two metrics I selected as my invariant metrics were the number of cookies and the click-through-probability. I choose the number of unique cookies to view the course overview page because the unit of diversion for the experiment is cookies, and there is no reason to expect that the number of cookies to view the course overview page should be statistically different between the control and the experiment group. The subjects will not be asked about the number of hours they plan to devote to the course per week until they have clicked on the "start free trial" button, which must occur after they visit the course overview page. The click-through-probability, or the number of unique cookies to click the "start free trial" button divided by the number of unique cookies to view the course overview page should not be statistically different between the control and the experiment group for the same reason that the experiment does not occur until after the cookie has selected the "start free trial" button.

I did not choose number of user-ids, number of clicks, gross conversion, retention, or net conversion as an invariant metric. Number of user-ids, or the number of users who enroll in the free trial, would be expected to differ between the control and the experimental group as users are presented with the question about the number of hours they are expecting to devote to the class before they begin the trial. Moreover, the unit of diversion for the experiment is unique cookies, not user-ids, and using a different unit of analysis and unit of diversion will increase the standard error of the experiment. Number of clicks, or the number of unique cookies to click the "start free trial" button could be used as an invariant metric because the click would occur before the experimental question is shown to the cookie. However, because I had already selected the click-through-probability as an invariant, it would be redundant to also use the number of clicks. The gross conversion, or the number of user-ids to complete checkout and enroll in the free trial divided by the number of unique cookies to click the "start free trial" button would be expected to differ between the control and the experiment group. I would expect that this number should statistically decrease if the experiment is successful because it would potentially show that users presented with the number of hours questions are not enrolling based on seeing the message about the number of hours they should plan on devoting to the course. The retention, or the number of user-ids to complete checkout and enroll in the free trial divided by number of unique cookies to click the "start free trial" button, is problematic as an invariant, because again, this metric could potentially vary between the control and experimental group. If the experiment is successful, this metric will not be statistically decrease between the experiment and control group, but it needs to be verified experimentally. Finally, net conversion, or the number of user-ids to remain enrolled past the 14-day boundary divided by the number of unique cookies to click the "start free trial" button, is a metric that could show a statistically significant difference between the experimental and the control group. Subjects in the experiment will see the message

after clicking on the button, which could subsequently affect the number of users who end up paying for the course.

Metric Choice: Evaluation Metrics

The two evaluation metrics I choose for the experiment were the gross conversion and the net conversion. I selected gross conversion, because if the experiment is successful, the gross conversion should demonstrate a statistically significant decrease. I would expect that users in the experiment who encounter the number of hours question and are not prepared to devote more than 5 hours/week to the course will be not as likely to complete checkout, therefore reducing the gross conversion metric for the experiment group. Net conversion is a suitable evaluation metric because it will serve as a check to ensure that the experiment does not decrease the number of students who eventually make a payment after starting the free trial.

The number of cookies, click-through-probability, and number of clicks were not selected as an evaluation because they would not be expected to vary between the control and experiment group. The number of user-ids to enroll in the free trial could be used as an evaluation metric, but a better choice would be a rate that is normalized by the number of user-ids to click on the "start free trial" button. Gross conversion already measures this rate, so the number of user-ids would be a redundant evaluation metric and more metrics increases the chance of a false positive. Finally, the retention would also be an acceptable evaluation metric. However, after completing some of the exercises further on in the experiment design, measuring the retention to an acceptable confidence level required an unfeasible number of pageviews. The large number of pageviews subsequently resulted in an experiment too lengthy to run on the Udacity website.

Expected Results for Success

For the experiment to be successful, both invariant metrics should not demonstrate a statistically significant difference between the control and the experiment group. This will be determined by computing a confidence interval for the invariant metrics and determining if the measure metric falls within this confidence interval. If it does, then the invariant metrics do not significantly vary and the experiment passes on the first sanity check. Furthermore, for the experiment to be green-lighted, there would need to be a statistically significant decrease in the gross conversion rate, and no statistically significant decrease in the net conversion between the control and the experiment group. The experiment is designed to dissuade students from beginning the free trial if they are not prepared to devote a minimum of five hours/week to the course and to divert them to access the course materials for free instead. Therefore, I would expect that in the experiment group, there will be some users who select the "start free trial" button, see the suggested hours message, and do not choose to complete checkout and enroll in the free trial. This would reduce the gross conversion rate as compared to the control group. In order for the experiment to be practically significant, this reduction must be at least 1% (dmin = 0.01). The net conversion should not decrease between the experiment and the control group if the experiment has the desired outcome. The users who see the message who already would have completed the course will not be affected by the message, and the users who see the message and normally would have canceled the free trial out of frustration should be diverted to access the

course materials for free rather than enrolling and then leaving the trial. In order for this metric to have a practically significant difference, the deviation between the control and the experiment group should be greater than 0.75% (dmin = 0.0075). If both invariant metrics do not demonstrate a statistically significant difference, the gross conversion demonstrates a statistically significant decrease, and the net conversion does not show a statistically significant decrease, then the experiment can be ruled a success and the change implemented.

Standard Deviation

Gross Conversion: 0.0202

Net Conversion: 0.0156

I would expect that the empirical variability and the analytic variability would be in agreement because the experiment has independent sampling. For both of the evaluation metrics, gross conversion and net conversion, the denominator, or the unit of analysis, is the number of cookies, which is also the unit of diversion used in the experiment. For this experiment, because the unit of diversion is equal to the unit of analysis, the sampling is independent and the analytic method can be used to estimate the experimental variability. At this point in the experiment, we do not know the distribution of the samples, but the analytic method is fairly robust to distribution and therefore the use of the analytic variability as an estimate of the variability is valid.

Sizing

Number of Samples vs. Power

Number of pageviews: 685,275

Duration vs Exposure

Fraction of Traffic Diverted: 0.6

Length of Experiment: 29 days

I choose to divert 60% of Udacity's course page traffic into the experiment. I felt this was the optimal compromise between a short duration and limiting exposure. Based on the experiment description, I thought there was a low risk to Udacity's users, which justified diverting more than half of traffic. Moreover, the choice of 60% meant that the experiment could be carried out in about a month, which would be a reasonable length of time to run the experiment while allowing time for data analysis. The two primary questions regarding risk to users that must be considered are whether or not the experiment can harm anyone, and whether or not the experiment collects sensitive information. In response to the first query, there is no certainly no chance of users being harmed physically, and users will also not see any negative effects on their studies because of the experiment. There is little risk to users as the student's learning experience is not significantly altered, and ultimately, the users still have the choice to enroll in the class even after seeing the message regarding the number of hours required per week to be successful. Although some students might be discouraged from enrolling in the free trial, these are students who likely would not be successful in the course in the first place, and they would be better off accessing

the course materials for free before making the free trial commitment. The second question, of whether or not sensitive information will be used, is also no because the experiment is only tracking whether or not the user-id makes a payment, and none of the details of the payment itself. Based on these answers, this is a minimal risk experiment, and directing up to 100% of the traffic would be justified. I will stick with 60% though because that results in a manageable length of experiment and will also allow for Udacity to monitor whether or not the experiment somehow affects non-experiment traffic.

Experiment Analysis

Sanity Checks

Number of Cookies: Confidence Interval = [0.4988, 0.5012], Observed Value = 0.5006, Pass

Click-through-Probability on “Start free trial” button: [-0.0013, 0.0013], Observed Value = 0. 0001, Pass

Both invariant metrics do not statistically differ between the control and the experiment group. Thus, the sanity check for the experiment passes and the analysis can proceed.

Result Analysis

Effect Size Tests

Gross Conversion: Confidence Interval = [-0.0291, -0.0120], STATISTICALLY and PRACTICALLY SIGNIFICANT

Net Conversion: Confidence Interval = [-0.0116, 0.0019], NOT STATISTICALLY and NOT PRACTICALLY SIGNIFICANT

Sign Tests

Gross Conversion: 19/23 successes, two-tailed p-value (probability of occurring by chance) = 0.0026, STATISTICALLY SIGNIFICANT

Net Conversion: 13/23 successes, two-tailed p-value = 0.6776, NOT STATISTICALLY SIGNIFICANT

Summary

After initially using the Bonferroni Correction, I decided against it for a re-analysis of the data based on feedback I received. I had used the Bonferroni correction in order to reduce the family wise error rate, or the probability of observing one or more false positives (Type I errors) when performing multiple hypothesis tests (using several metrics) on the same dataset. I made this decision because it did not result in an unfeasible number of pageviews, and it reduced the chance that I would record a false positive when performing the analysis on the data. However, the issue with the Bonferroni Correction is that it can be conservative, meaning that it increases the chance of recording a false negative (type II error). In order to make a decision about the tradeoffs of using the Bonferroni Correction, I had to consider the style of the experiment being

run. The principal question was whether a launch of the change could be triggered by ANY metric recording a statistically and practically significant result, or whether ALL metrics would have to record a statistically and practically significant in order for the experimental condition to be implemented. In the first case, with ANY metric triggering the change, false positives take on the greatest impact because even a single false positive would cause the change to be falsely implemented. In this case, the Bonferroni Correction should be used to minimize the change of a false positive. If ALL metrics must show a result in order to launch, false negatives take on a larger importance because even a single false negative would falsely prevent the experimental change from being implemented. In this case, the Bonferroni Correction should not be used because it increases the probability of a false negative. For this experiment, I defined success as gross conversion recording a significant result and net conversion not recording a significant result and therefore the experiment falls under an ALL situation. The Bonferroni Correction should not be used because a single false negative would prevent launch. Therefore, I revised my answers to reflect my updated decision to not use the Bonferroni Correction for the experiment.

There were no discrepancies between the effect size testing and the sign testing.

Effect Size Diagrams

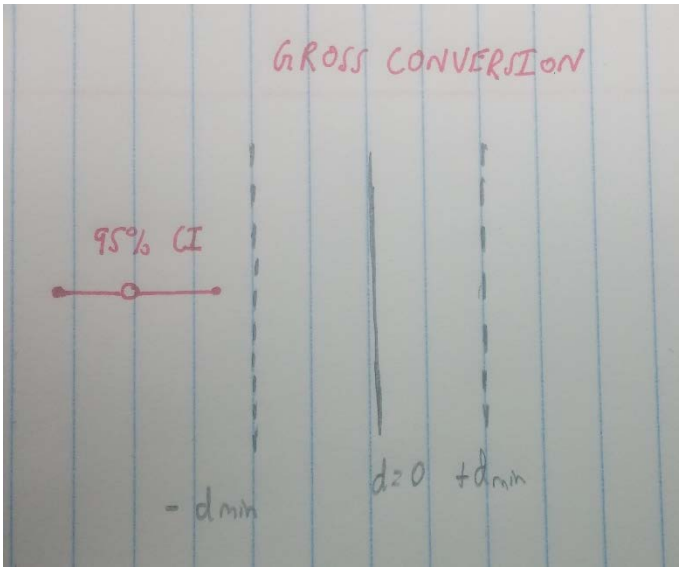

Figure 1: Gross Conversion Effect Size

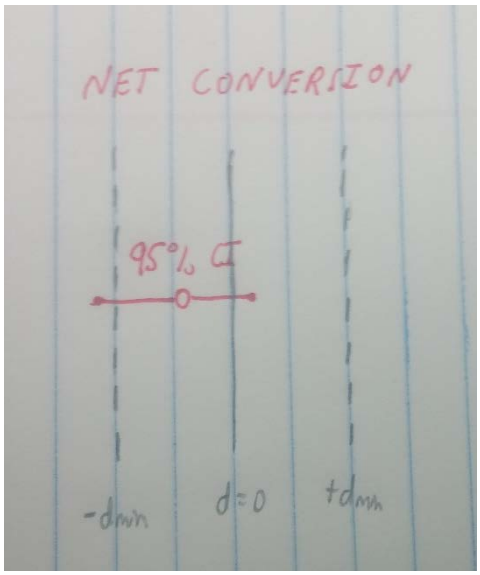

Figure 2: Net Conversion Effect Size

Recommendation

I would recommend that Udacity perform further testing before any change is implemented. Based on the experimental results, asking users about the number of hours they expect to devote to the course per week and then warning the users that five hours per week is the minimum recommended time commitment for completion of the course, did achieve the objectives of the experiment with a minor qualification. The gross conversion rate, or the number of unique user-ids that enrolled in the free trial divided by the number of unique cookies that clicked on the

“Start free trial” button showed a statistically significant decrease between the experimental and the control group (see Figure 1). My conclusion from this is that there were a significant number of students who clicked on the button, entered a weekly hour figure less than the recommended amount, were warned about the facts regarding course completion, and consequently choose not to enroll in the free trial. These students were probably the most likely to drop out of the free trial due to frustration, so diverting them before they could start the trial is a good idea. The other evaluation metric also met the standards of success established before the experimental analysis. The net conversion did not show a statistically or practically significant decrease. However, examining the 95% confidence interval for the net conversion, I see that the bottom limit of the interval falls below the negative of the minimum change required for practical significance (see Figure 2 below). The bottom limit of the CI was -0.0116, which is beyond the -0.0075 which forms the practical significance level. Although not all of the CI is below this value, at least some of it is which suggests that the experimental group might have recorded a decrease in the number of users who go on to pay for the class after clicking the “start free trial” button. This could show that the “free trial screener” does in fact decrease the number of users who go on to enroll and pay for the course, which would run counter to the objectives of the change. Therefore, I would recommend that Udacity perform further testing before implementing the “start free trial” screener for all website users. The experiment suggests that at least one of the objectives is met, but it also shows that there is a possibility that the other evaluation metric does not pass, and another round of testing should be run to resolve any inconsistencies.

Follow-Up Experiment

A potential follow-up to decrease the number of frustrated students who cancel early in the course (after enrolling in the free trial) would be to offer students who do not pass the first projection submission in the first two trials a free coaching lesson. This coaching lesson would guide the students through the project as well as the general overall process of improving a submission in order to meet the requirements on the rubric. The lesson would be offered to the students after they have failed the project for a second time and would be presented along with the second project review. Passing a project can be a huge confidence boost, but the first project can be difficult to figure out on one’s own. I know that I was frustrated by the first real project (Investigating a Statistical Phenomena) even though I did not find the material overly difficult. The challenge was more in learning the style of Udacity’s course and figuring out how to satisfy all the project requirements. My hypothesis is that offering students a free coaching session would decrease the number of students who begin the free trial and then drop out because they cannot pass the first project submission. Moreover, this would show students one of the resources available and could potentially increase the number of coaching sessions purchased throughout the course by the student if they find it a valuable experience. An invariant metric would be the number of students who fail the project twice. As the users are not presented with the option for the free coaching lesson until after they have failed the project for the second time and thus the number of users failing the project twice should be equivalent in the control and experimental group. Another invariant metric to use as a sanity check would be the average number of specifications for each user that need to be corrected on the second project review.

Again, this would be determined before the experimental change is shown to the users and this metric should not show a statistically significant change between experiment and control groups.

I would use a metric of diversion of user-id. This is because once a user begins the free trial, they are tracked by user-id and I want to be tracking the effect of the change for users across the entire course. I would use the metric of retention from the free trial screener experiment which is the number of user-ids to remain enrolled past the 14-day boundary (completing at least one payment) divided by the number of user-ids to complete checkout. I would use a practical significance level of 1%, and for the experiment to be a success, I would need to see this metric increase at both a statistically and practically significant level. An additional metric I would use would be the project 1 successful completion rate after two failed submissions. The formal definition would be the number of users to complete the first project after failing twice divided by the total number of users to fail the first project twice. The experiment would be successful if this rate was increased at a practical significance level of 1%. Using both of these metrics would let me assess whether the free coaching lesson increases the number of students who go on to pay after joining the free trial, as well as the direct effect of the coaching lesson on the students rate of passing the first project. The experimental group, or the group shown the video after failing the first project twice, should see an increase in both metrics in order for the experiment to be deemed a success. The experiment might require a long time frame because the number of users who fail the first project in the first two submissions is likely low. Therefore, variations on the experiment could be run with users who fail the first submission a single time, or the free coaching lesson could be shown to all students before they even submit the project.

Resources

http://www.evanmiller.org/ab-testing/t-test.html

http://staweb.sta.cathedral.org/departments/math/mhansen/public_html/23stat/handouts/normbino.htm

http://www.aaos.org/AAOSNow/2012/Apr/research/research7/?ssopc=1

https://docs.google.com/document/d/16OX2KDSHI9mSCriyGIATpRGscIW2JmByMd0ITqKYvNg/edit

https://docs.google.com/document/u/1/d/1aCquhIqsUApgsxQ8-SQBAigFDcfWVVohLEXcV6jWbdI/pub?embedded=True

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

https://graphpad.com/quickcalcs/binomial1.cfm

http://onlinelibrary.wiley.com/doi/10.1111/opo.12131/full

https://discussions.udacity.com/t/bonferroni-correction/201344