Linh-Tam Trinh
January 31, 2017

Project 7: A/B Testing

Experiment Design

Metric Choice

For invariant metrics, I used the Number of Cookies, Number of Clicks, and Click-through-probability. These were selected as invariant metrics because they were all presented to users before the free trial screener and help gauge if the traffic to both the experiment and control group were comparable. They would not be good evaluation metrics because they are presented before the free trial screener and would not measure the effect of the free trial screener.

For evaluation metrics, I used Gross Conversion (number of enrollment/number of clicks) and Net Conversion (number of payment/number of clicks). These measure the effect of having the free trial screener and help evaluate whether or not having the free trial screener would affect users enrolling and remaining enrolled long enough to make payment. They would not be useful as invariant metrics as they measure the effect and are expected to be different between the control and experiment groups.

For Gross Conversion, a difference less than the lower bound (-0.0291) or greater than the upper bound (-0.0120) would be statistically significant. The practical significance that would make this experiment substantive is set at dmin=0.01, so the difference would need to be less than -0.01. For Net Conversion, a difference less than the lower bound (-0.0116) or greater than the upper bound (0.0019) would be statistically significant. Since we are looking for the Net Conversion to not decrease, with the practical significance at dmin=0.0075, the difference would need to be higher than -0.0075.

A couple metrics were not used as invariant or evaluation metrics. The Number of User Ids that enroll in free trial was not used as an invariant metric or evaluation metric because it was not equalized by the Number of Clicks or Number of Cookies and could be biased towards whichever group has more traffic. Retention was also not used as an invariant or evaluation metric. Since it measures the effect of the free trial screener and is expected to be different for the experiment and control groups, it would not be a useful invariant metric. Although it could be useful as an evaluation metric, it is not used because the number of trials that it requires would be too many and would prolong the duration of the experiment longer than desired for A/B testing.

Measuring Standard Deviation

The standard deviation for Gross Conversion is 0.0202 and the standard deviation for Net Conversion is 0.0156. I expect the analytic estimate to be comparable to the empirical variability for both metrics since the unit of diversion and the unit of analysis are the same. The unit of diversion is cookies, since cookies are used to divide the participants between the control and experiment group. The unit of analysis for both Gross Conversion and Net Conversion is also cookies, since the denominator for both is the number of cookies to view the page.

Sizing

Number of Samples vs. Power

At an alpha = 0.05 and beta = 0.20, and without using the Bonferroni correction, I would need 685,325 page views to power my experiment appropriately.

Duration vs. Exposure

I would assess this experiment as very low risk in regard to the principles of IRB. The risk to the participants in this experiment would be minimal as it does not harm them physically, psychologically, emotionally, socially, or economically. There is no sensitive information (medical, financial, etc.) and users have the option not to use Udacity. In addition, the experiment would help Udacity benefit by improving its product while preventing users from wasting time from cancelling. Since this experiment is minimal risk and requires a significant amount of pageviews, I would divert 100% of traffic to it, lasting approximately 18 days.

Experiment Analysis

Sanity Checks

All of the invariant metrics had an observed difference within the 95% confidence interval, passing the sanity checks and indicating that the experimental setup is reasonable. The Number of Cookies had a lower bound of 0.4988 and upper bound of 0.5012 and a passing observed bound of 0.5006. The Number of Clicks had a lower bound of 0.4959 and upper bound of 0.5041 and a passing observed bound of 0.5005. The Click-through-probability had a lower bound of 0.0812 and upper bound of 0.0830 and a passing observed bound of 0.08213.

Result Analysis

Effect Size Tests

The Gross Conversion has a 95% confidence interval around the difference between the experiment control group with a lower bound of -0.0291 and upper bound of -0.0120, being both statistically and practically significant. The Net Conversion has a lower bound of -0.0116 and upper bound of 0.0019, being both statistically and practically not significant.

Sign Tests

Running a sign test for Gross Conversion resulted in a statistically significant p-value of 0.0026. However, the sign test did not result in a statistically significant value for Net Conversion with p-value equals 0.6776.

Summary

I decided not to use the Bonferroni correction because it would not be relevant. In a multiple metric analysis, the Bonferroni correction can be helpful in reducing type I errors (false positives) if the experiment is looking for *any* of the metric to meet expectation. In this experiment, since *all* of the metrics need to meet expectations, the use of multiple metrics actually makes the requirement more stringent, thus the Bonferroni correction would not be necessary.

There was no discrepancy between the effect size tests and sign tests, the Gross Conversion was statistically and practically significant in the effect size test and was also statistically significant in the p-value sign test. Meanwhile, the Net Conversion was not statistically and practically significant in the effect size tests and was also not statistically significant in the p-value sign test.

Recommendation

I would recommend to run additional tests. The Gross Conversion decreased as expected, with the effect size test being statistically and practically significant and the sign test also being statistically significant. This means that the number of people who enrolled for the free trial decreased with the free-trial screener. The issue is with the Net Conversion. Although the effect size test was not statistically or practically significant and the sign test was also not statistically significant. Meaning that there was no significant change in the number of people who made a payment given the free-trial screener. It is important to note that the confidence interval (-0.0116 to 0.0019) includes the negative of the practical significance boundary, meaning that the experiment could possibly make the Net Conversion decrease. The result shows that the free-trial screener is effective at reducing the number of people who sign up and quit, however, it could possibly be at the cost of some people who stay and make payment. To determine if there is a decrease in Net Conversion, I would recommend to run additional tests with more power.

Follow-Up Experiment

I think that a main reason why students cancel early in the course is because they do not know how to complete the projects and do not know how to get help, so they become frustrated and cancel. To help with this problem, I think that Udacity should make a cheat sheet/study guide on how to complete a very similar project with step-by-step guides and explanations that is repeatable for the real projects. Students do not have to use this guide, but it is an option that they could use to help them through confusing problems and ensure that they are able to complete their projects if they are committed to it.

For the experiment, nothing would change for the control group, meanwhile for the experiment group, there would be a link to the Cheat Sheet on the Project Overview page. I think that the people given the option to use the Cheat Sheet will be more likely to complete the project. Although they do not have to use the Cheat Sheet, having it there to reference when they are

struggling could be very helpful. An invariant metric would be the number of User-Ids to view the Project Overview page (before clicking on the Cheat Sheet link). This would serve as an invariant metric by helping us gauge that there is a comparable amount of people who reach the Project Overview page for both the experiment and control group. An evaluation metric would be the Completion-to-View Percentage (the number of User-Ids to complete the project divided by the number of User-Ids to view the Project Overview page). This would help us compare whether the people in the experiment group, with the option to use the Cheat Sheet would be more likely to complete the project. The unit of diversion in this experiment would be the number of User-Ids. My hypothesis is that the Complete-to-View Percentage in the experiment group would be higher than in the control group.