A/B Testing Final Project

by SAQIB ALI

Choice of Metrics

List which metrics you will use as invariant metrics and evaluation metrics here.
Invariant metrics: Number of cookies, Number of clicks
Evaluation metrics: Gross conversion, Retention, Net conversion

For each metric, explain both why you did or did not use it as an invariant metric and why you did or did not use it as an evaluation metric. Also, state what results you will look for in your evaluation metrics in order to launch the experiment.

1. Number of cookies: Number of unique cookies to view the course overview page. The visits happen before the user sees the experiment, therefore independent of the experiment. So it is invariant metric.
2. Number of user-ids: Not a good invariant metric because the number of users who enroll in the free trial is dependent on the experiment. Not a good evaluation metric because the number of visitors may differ between the experiment and control groups, which would skew the results. The number of user-ids or enrolled users can fluctuate a lot with respect to the number of start free trial clicks on a given day, and thus not a good proxy for this experiment. Since gross conversion can indicate a relative difference between the number of enrollments, we rather not use it.
3. Number of clicks: The number of clicks is the number of unique cookies to click on the" start free trial" button. This is invariant metric because the clicks happen before the user sees the experiment, and are thus independent from it.
4. Click-through-probability: Click-through-probability is the number of unique cookies to click on the 'start free trial' button divided by the number of unique cookies to view the course overview page. Good invariant metric because the clicks happen before the user sees the experiment, and are thus independent from it. We know that the number of cookies and number of clicks are already sufficient to use as invariant metric. Extra metic is as critical as the other two.
5. Gross conversion: It is a good evaluation metric because it is directly dependent on the effect of the experiment and will allow us to show whether we managed to decrease the cost of enrollments that aren't likely to become paying customers.
6. Retention: As the number of users who enroll in the free trial is dependent on the experiment, so it is not a good invariant metric. But it is a good evaluation metric because it is directly dependent on the effect of the experiment, and also shows positive financial outcome of the change.
7. Net conversion: Net conversion is the number of users to remain enrolled past the 14-days boundary (and thus make at least one payment), divided by the number of unique cookies to click on the "start free trial" button. It is not an invariant metric because the number of users who enroll in the free trial is dependent on the experiment. It is a good evaluation metric because it is directly dependent on the effect of the experiment, and also shows positive financial outcome of the change.

At the end of the experiment, if the evaluation metric is practically significant and better than the control group at the end of the experiment, we can launch the new feature. We will focus on Gross conversion and Net Conversion because using Retention as evaluation metric will require too many more pageviews (4,741,212).

We expect gross conversion will decrease practically significant, which will indicate whether the cost will be lower by introducing the new screener; while net conversion will not decrease statistically significance, which will indicate the screener whether or not affect the revenues.

Measuring Standard Deviation

List the standard deviation of each of your evaluation metrics.
Gross conversion: 0.0202
Net conversion: 0.0156
Retention: 0.0549

For each of your evaluation metrics, indicate whether you think the analytic estimate would be comparable to the empirical variability, or whether you expect them to be different (in which case it might be worth doing an empirical estimate if there is time). Briefly give your reasoning in each case.

Gross conversion and net conversion both have the number of cookies as their denominator, which is also our unit of diversion. Here, the unit of diversion is equal to unit of analysis, which indicates that the analytical estimate would be comparable to the empirical variability. So we can proceed using an analytical estimate of the variance.

Sizing

Number of Samples vs. Power

Indicate whether you will use the Bonferroni correction during your analysis phase, and give the number of pageviews you will need to power you experiment appropriately
The Bonferroni correction is designed to limit the risk of Type I errors in multiple comparisons. It can be definitely used in cases where multiple independent tests performed simultaneously, and to make a decision, we expect at least one of them to be statistically significant. To launch the experiment, We expect both gross conversion and net conversion to be significant. The metrics in the test have high correlation (covariant) and the Bonferroni correction will be too conservative to it.

To adequately power experiment, we will need 685,325 pageviews, as we won't be using Retention as our evaluation metric.

Metric	*Pageviews*
Gross Conversion	646,450
Net Conversion	685,325
Retention	4,741,212

Indicate what fraction of traffic you would divert to this experiment and, given this, how many days you would need to run the experiment.
With daily traffic of 40000, we will direct 70% of traffic (28000) to the experiment, which means it would take us approximately 25 for the experiment.

Give your reasoning for the fraction you chose to divert. How risky do you think this experiment would be for Udacity?
The experiment is not very risky as it is not going to affect the normal operations. The implementation is also very simple, so there are very few chances of introducing a bug to the application itself. But we will not suggest 100% traffic to be diverted to experiment page, in case there is an issue, only some fraction of students is affected.

Sanity Checks

For each of your invariant metrics, give the 95% confidence interval for the value you expect to observe, the actual observed value, and whether the metric passes your sanity check.

Metric	*Confidence interval*	*Observed*	*Outcome*
Number of cookies	[.4988, .5012]	0.5006	PASS
Number of clicks	[.4959, .5041]	0.5005	PASS
Click-through-probability	[.0812, .0830]	0.0822	PASS

Result Analysis

For each of your evaluation metrics, give a 95% confidence interval around the difference between the experiment and control groups. Indicate whether each metric is statistically and practically significant.
A Z-test indicated that the gross conversion was statistically and practically significantly lower in the experiment group than in the control group, dmin = 0.01, 95%CI[−0.0291, −0.0120].

A Z-test indicated that the net conversion was neither statistically nor practically significantly different between the experiment and control groups, dmin = 0.0075, 95%CI[−0.0116, 0.0019].

Metric	*Confidence interval*	*Statistically Significant*	*Practically Significant*
Gross Conversion	[-.0291, -.0120]	YES	YES
Net Conversion	[-.0116, .0019]	NO	NO

Sign Tests
For each of your evaluation metrics, do a sign test using the day-by-day data, and report the p-value of the sign test and whether the result is statistically significant.

Metric	*p-value*	*Statistically Significant*
Gross Conversion	0.0026	YES
Net Conversion	0.6776	NO

Summary

State whether you used the Bonferroni correction, and explain why or why not. If there are any discrepancies between the effect size hypothesis tests and the sign tests, describe the discrepancy and why you think it arose.

I did not use a Bonferroni correction because we are only testing one variation. It might be useful to apply the Bonferroni correction if we decide to do post-test segmentation on the results, for example based on browser type or countries of origin.

Recommendation

Make a recommendation and briefly describe your reasoning.

The evaluation metrics we focused on were Gross Conversion and Net Conversion. Gross conversion turned out to be negative and practically significant which meant we lower our costs by discouraging trial signups that are unlikely to convert. But Net Conversion ended up being statistically and practically insignificant and confidence interval includes negative numbers. Therefore, there is a risk that the introduction of the new feature may lead to a decrease in revenue.

Follow-Up Experiment

Give a high-level description of the follow up experiment you would run, what your hypothesis would be, what metrics you would want to measure, what your unit of diversion would be, and your reasoning for these choices.

I have been doing a lot of research about the course provided related to Data Analysis and going through the free courses. But one day I saw an offer, Rio Olympics 2016 celebratory 17% discount worth USD 32 a month, which made me sign up as paying student for Nanodegree. Although there is already a discount of 50% by returning half fee if program finished before a set duration. But this new discount made the deal even sweeter.

The proposed discount will show up on top of course training page, for half of the enrolled students, diverted by ID. The hypothesis is that by providing this additional discount of lowering the monthly cost, we will see increase in paying sign ups as this feature will be potentially compelling to users who are already determined to take the course and ready to jump in directly.

I would use user-ID as the unit of diversion, as the change would only be visible to users whose are enrolled in a course. I would use user-ID as an invariant metric, and the number of payments divided by the number of user-IDs as an evaluation metric.

Resources

1. https://vwo.com/ab-testing/
2. http://abtestguide.com/calc/