Project 7
Design an A/B Test

Marc Wu

25. January 2017

Experiment Design

Metric Choice

Invariant Metrics

Invariant metrics are expected not to be significantly different between control and experiment groups. They serve as a sanity check to ensure that both groups are comparable and the experiment was done properly. Generally, there are two types of invariant metrics: population sizing metrics and any other metrics we do not expect to change.

Number of cookies (number of unique cookies to view the course overview page)
Cookies are counted before the student clicks on “Start free trial”, therefore this metric should be invariant across both groups. In addition to that, the cookie is our unit of diversion, which is per se a good invariant metric with regards to the population size.

Number of clicks (number of unique cookies to click the "Start free trial" button)
We expect this metric to be equivalent in both groups, because the tested change happens after clicking the “Start free trial” button. In other words, the user experience is the same up to the point when this metric is recorded.

Click-through-probability (number of unique cookies to click the "Start free trial" button divided by number of unique cookies to view the course overview page)
This metric is calculated via the two metrics discussed before, hence it should be invariant as well.

Evaluation Metrics

Evaluation metrics are used to justify business decisions, which in our case is the launch of the change (or not). We can put evaluation metrics into two categories. The ones we hope to observe a statistically significant difference with a minimum detectable effect (practical significance level d_{min}). The second category comprises evaluation metrics we hope not to get worse with the implemented change.

Udacity wants to achieve two goals with the free trial screener:

1. Reduce the number of frustrated students who left the free trial because they didn't have enough time
2. No significant reduction in the number of students to continue past the free trial

Gross conversion (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; d_{min}= 0.01)
We will use this metric to measure the first goal, because the screener should prevent students with not enough time to enroll. This will result in less frustrated students, but also in less total number of free trial enrollments, therefore decreasing the overall gross conversion.

Net conversion (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; d_{min}= 0.0075)
This metric covers our second goal. If we have decreased gross conversion (less people enrolling), we need (relatively) more students to remain enrolled past the free trial to have no reduction in net conversion. We hope that this is the case, because frustrated students would not have stayed enrolled anyways.

Expected Results
In order to recommend the launch of the free trial screener, we want to observe a statistical significant decrease in gross conversion with an effect size of at least 1% together with no statistical significant reduction in net conversion.

Ignored Metrics
Number of user-ids (number of users who enroll in the free trial)
Since this metric is recorded after being exposed to the change, it is not an adequate invariant metric. It also not used as an evaluation metric, because it is a count which is just an absolute value. An increase in number of user-ids does not tell us if it is due to higher conversion rates or because there are more people visiting the page. Therefore it is more appropriate to have a rate or probability as an evaluation metric.

Retention (number of user-ids to remain enrolled past the 14-day boundary divided by number of user-ids to complete checkout)
Initially, I picked retention as an evaluation metric, because it covers both goals of Udacity. An increase of this metric should be due to less people completing the checkout while having no reduction in the number of students making a payment. Unfortunately, in later stages of my analysis it turned out that the required duration of the experiment would have been too long. Therefore, I removed it as an evaluation metric.
Retention can not be used as an invariant metric for the same reason which was given for the number of user-ids metric.

Measuring Standard Deviation

Evaluation Metric	Analytical Standard Deviation
Gross Conversion	0.0202
Net Conversion	0.0156

If the unit of diversion is the same as the unit of analysis, then the empirical and analytical variabilities are likely to match. Since cookie is the unit of diversion and it is used in both metrics, I assume that the analytic estimate is comparable to the empirical standard deviation.

Sizing

Number of Samples vs. Power

No Bonferroni correction is used in the analysis phase.
I used http://www.evanmiller.org/ab-testing/sample-size.html to calculate the required page views for each evaluation metric. A significance level of $\alpha = 5\%$ and a statistical power of $1-\beta = 80\%$ were used:

Metric	Baseline Conversion Rate	d_{min}	Required Sample Size per group	Required page views per group	Total required page views
Gross Conversion	20.625%	1%	25835	322937	645874
Net Conversion	10.93125%	0.75%	27413	342662	685324
Retention	53%	1%	39115	2370606	4741212

I added the numbers for retention to show that it needs approximately 7 times more page views than the other two metrics.

In the end we need 685324 page views to power the experiment appropriately.

Duration vs. Exposure

We need to run the experiment for more than 17 days if we divert all the traffic (40000 page views per day) to our experiment.

Choosing a diversion rate is based on several factors:

- Risk for the business: the more risky the experiment is, the less fraction of traffic we want to divert and vice versa. In addition to that, if the risk is high we also want to keep duration short for minimal business impact. Risk could involve the uncertainty of the feature working correctly, or how the users will react to the change.
  There is a possibility, that the number of payments would decrease, which may be unacceptable depending on the magnitude. However, I rate the possibility of a steep decline quite unlikely, because the change is just an additional hint. Considering this I conclude that the experiment is not particularly risky for Udacity.
- Risk for the participants: an experiment can pose risks for the participating subjects. Depending on the severity, we need to keep exposure low for ethical reasons. In our experiment there is minimal danger that the student would suffer from physical, psychological, or financial harm, because the screener is only a subtle change. Moreover, no sensitive nor additional data is being gathered.
- Other experiments: we need to consider other experiments if they are running simultaneously. If it is the case, we can not redirect all traffic to one experiment only and have to make sure that each experiment receives its fair share. There are no other experiments running in our case, therefore we can neglect this factor.
- Duration: in general we want to keep duration short, so we can run other experiments subsequently and make timely business decisions. However, too short durations can lead to unwanted biases, e.g. holiday versus regular season, weekday versus weekend.

After assessing these factors I choose a diversion of 65% resulting in 27 days required to run the experiment.

Experiment Analysis

Sanity Checks

I computed a 95% confidence interval for the difference of the two means (experiment and control group), to see if both groups are comparable. The results are summarized in the following table:

Invariant Metric	Lower bound	Upper bound	Observed	Passes
Number of cookies	0.4989	0.5011	0.5006	Yes
Number of clicks on “Start free trial”	0.4959	0.5041	0.5004	Yes
Click-through-probability on “Start free trial”	-0.0012	0.0013	0.0000	Yes

Result Analysis

Effect Size Tests

The 95% confidence intervals for the difference between the experiment and control groups are given in the table below. If the confidence interval does not include 0 then the metric is statistically significant. We have practical significance if d_{min} is outside the confidence interval.

Evaluation Metric	d_{min}	Lower bound	Upper bound	Statistical significance	Practical significance
Gross conversion	0.01	-0.0291	-0.0119	Yes	Yes
Net conversion	0.0075	-0.0116	0.0018	No	No

Sign Tests

Metric	p-value	Statistical significance
Gross conversion	0.0026	Yes
Net conversion	0.6776	No

Summary

No Bonferroni correction was used, because we only launch our change if every evaluation metric shows statistical significance. The correction is needed when statistical significant difference for only some of the metrics are sufficient for a launch decision. The reason is, with increasing number of evaluation metrics we increase the chance of a false positive, i.e. we reject the null hypothesis although it is true.

Both effect size and sign tests are consistent with their results: gross conversion shows statistical significance while net conversion does not at a 95% confidence level. The lack of discrepancy increases the trust in our findings and will support our following recommendation.

Recommendation

Gross conversion showed a statistically significant decrease of 2%, well above the required practical significance level of 1%. We can conclude that the screener reduced enrollments successfully, which was one of the goals.

Unfortunately, the results for net conversion are not clear, because they are not statistical significant. We can not say whether net conversion has increased, decreased, or if it did not change at all, because the corresponding confidence interval spans negative and positive numbers.

Therefore, I strongly recommend not to launch the change, because it could lead to decreased net conversion numbers, posing a harmful business risk. Instead I suggest to investigate net conversion more further.

Follow-Up Experiment

Our A/B test was all about the reduction of students with not enough time which would have canceled anyway. Next, I would investigate the students who completed the checkout and think about ways how to increase their rate to stay enrolled after the free trial.

One major advantage of the paid version of a course are the moderated forums. This benefit may not become obvious for the student during the free trial period. Therefore, I suggest to implement a message shown after the checkout, which asks the student to make at least 3 posts in the forums over the course of the free trial.

My expectation is that forum engagement increases overall commitment and motivation, reduces frustration because of quick and competent help, and gives you the sense to be part of an awesome community. This should lead to increased retention, which would be my evaluation metric of choice. Since students are tracked by user-id after enrolling, it would be the unit of diversion and thus an appropriate invariant metric.

Null hypothesis: asking the student to make at least 3 posts in the forums does not increase retention.
Alternative hypothesis: asking the student to make at least 3 posts in the forums increases retention.
Invariant metric: user-id
Evaluation metric: retention
Unit of diversion: user-id