A/B Testing

Experiment Design

Metric Choice

Invariant metrics verify whether the experimenter has divided the population appropriately into experimental and control groups. Once divided, the equivalence of the two groups is judged by the lack of variance between of each group's invariance metrics. Based on both the unit of diversion and the experiment's design, the invariance metrics 'check the sanity' of the assumptions made about equivalence of the control and experimental groups with empirical methods. Verifying the equivalence within a margin of error fulfills the task of homogenizing the population across groups, effectively reducing the probability of the study's results' owing to random chance.

I chose the number of cookies and the click-through-probability as invariance metrics. The number of cookies is an appropriate population sizing metric, as creating groups of equal size creates a more robust means of comparison. Because cookies are affected by a user's behavior, if the number of cookies differs significantly between the experimental and control groups, the experimenter can register that anomaly and explore its basis. Because users click the "Start Free Trial" button before the prompt is triggered, changes in the click-through-probability will not be affected by design changes in the experiment. The user IDs were not used as an invariant metric because cookies were the units of diversion. People could have multiple accounts, conditions under which the results would be skewed. On the other hand, user IDs are a suitable evaluation metric because they can represent the number of students who would continue past the free trial.

Evaluation metrics test the impact of changes to a website. They test the success of a business goal with respect to a change in the user's experience. The simplest outcome of an evaluation metric involves a business' profit or loss due to differences in the user's experience of the website.

The Udacity experiment hypothesized that adding a prompt would reduce dropout rates. The prompt, the thinking went, "might set clearer expectations for students upfront, thus reducing the number of frustrated students who left the free trial because they didn't have enough time...without significantly reducing the number of students to continue past the free trial and eventually complete the course.” We’re looking for metrics gauging the enrollment rate and payment rate in the control and experimental groups. Conversion and retention metrics, which measured either enrollment in the free trial or payment after the free trial, served to gauge the impact of the proposed changes. Gross conversion is the proxy for the rate of enrollment in the free trial, and net conversion is the rate of those who paid after 14 days of free trial. The experiment’s hypothesis is confirmed if net conversion doesn’t decrease and gross conversion decreases. That combination signals success in the experimental setting. Any other combination disconfirms the two hypotheses. Gross conversion tracks with the first statement above, and net conversion answers the second.

After testing was undertaken to determine the sample sizes necessary for a statistically significant reading of a given metric, retention was no longer included as an evaluation metric as it required hundreds of days of data collection in order to fully complete. The gross and net conversion rates were retained, requiring by contrast only a month’s-worth of data collection.

Measuring Standard Deviation

Analytical variability includes assumptions about the independence of the data, and differences in the empirical and analytic variability arise when the independence assumptions for the unit of analysis do not hold for those of the unit of diversion. If the unit of diversion is event-driven, independence assumptions are preserved. Every event is a random draw, and assumptions about the data’s independence assumptions should hold. A click-through-rate (CTR) metric illuminates the effect of these differences. If a cookie or a user-id is used as the denominator of a CTR, where the distinct user’s events can batch together, there exists potential for variance between the unit of analysis and the unit of diversion.   It is thus possible that the empirical and analytical variability diverge in such cases.

Gross conversion variability

.0202

Net Conversion

.0156

Empirical sample deviation should be comparable to the analytical deviation estimate for the gross and net conversion because the unit of diversion, a cookie, and the unit of analysis are the same.

Sizing

Number of Samples vs. Power

The significance of a test's result deriving from the results of a number of different tests is subject to at least two things. The strength of the interpretability of a collection of metrics relies on the sheer number of metrics, as well as each result's degree of independence from all of the other results. When a test relies on the agreement of multiple results, the significance of its findings degrades from that of any one of its composite tests due to the dilution of a single probability through the pooling of individual probabilities which exhibit at least some independence from one another. This introduces the multiple comparisons problem, which is a failure of statistical inference based on a specious understanding of the agreement of a number of component results as confirming the hypothesis of the whole argument.   This effect grows with the independence of the individual tests, and as more tests are included in the overall hypothesis test. Corrections, including the Bonferroni correction, are attempt to redress this problem by generally require a higher significance threshold for each test, thus compensating for the difficulties introduced by the multiple comparisons problem.

I chose not to use a Bonferroni correction for analysis. In order to launch the new feature, the experiment looks for a decrease in gross conversions, and no change in net conversions. Bonferroni corrections are used to reduce the risk of Type I errors, or the finding that the null hypothesis is rejected by chance alone. If the findings of every test in the analysis have to match to confirm a set of hypotheses, as is true of this experiment, a correction is not necessary to mitigate the risk that the rejection of any one metric undermines the findings of the metrics considered together, or a Type I error.

Additionally, web experiments must consider experimental risk. Experiments encounter risk when subjects are compromised by an experiment, be it through private or incriminating data released, or through harm done to the user's experience which negatively affects the business. Risk mitigation controls the degree to which these effects are constrained. Risk mitigation tactics include decreases in traffic diversion or data anonymization methods. This experiment, however, is a priori neither risky for users nor for the business. It avoids substantial changes in the UI, should not affect the business materially and doesn't require private data.

Duration is an experimental consideration as well. Experiments also shouldn't last too long, as they are expensive and prevent other experiments from being run. Had retention been a metric, the experiment's duration would have been too long (237 days) to be convenient. Otherwise, sufficient traffic is gathered within about 35 days for both gross and net conversion.

The web traffic will therefore not require risk mitigation and one step to decrease its duration.

Retention
n=4739879

Duration vs. Exposure

50% of traffic diverted with 237 days of experiment duration.

While 50% gives a sense of equivalence between the experimental and control group, this is too long a period of time to conduct an experiment for a quickly-moving business. Retention was therefore removed as an evaluation metric. With the new net retention metric and 50% of the traffic diverted, the duration of the experiment became 35 days, a manageable figure.

Experiment Analysis

Sanity Checks

Number of cookies-
CI: .4988-.5011
Observed: .5006, passed sanity check

Click-through-probability-
CI: -.0013-.0013
Observed: .0001, passed sanity check

Result Analysis

Effect Size Tests

Gross conversion dhat CI min: -0.02912335834
Gross conversion dhat CI max: -0.012

Both practically and statistically significant as dhat max does not cover zero, therefore this is a statistically significant decrease.

Net conversion dhat CI min: -0.0116
Net conversion dhat CI max: .0019

Neither practically nor statistically significant as the confidence interval includes the zero boundary.

Sign Tests

Gross conversion--.0026
Net Conversion--.6776
The gross conversion sign test was determined to be statistically significant, while the net conversion sign test was not.

Summary

I ran an analysis to determine the effect of a change to the Udacity course overview webpage. It was hypothesized that introducing a prompt would set the user's expectations such that both the rates of free trial dropout would reduce without reducing the number of students to complete the course. The experiment was undertaken by dividing Udacity's web traffic along the traffic for which the prompt occurred, or the experimental group, and the traffic for which the website behaved as it did originally, or the control group. Two invariant metrics, cookies and click-through-probability, were chosen to attest to the equivalence of both groups, and three evaluation metrics were originally chosen to monitor the changes between the control and experimental groups. It was determined that, in order to generate the sample size necessary to validate the significance of the retention metric, too much time was going to be devoted to data collection. Therefore, retention was dropped as an evaluation metric, leaving gross and net conversion as the remaining evaluation metrics. These metrics helped gauge the percentage of users who, having navigated to the webpage of interest, end up enrolling in a free trial or paying for a course.

I did not use the Bonferroni correction, as the result of every test needed to match expectations in order to launch the new feature. A high significance threshold for the most constrained sufficed to meet the significance threshold for a test of both metrics in conjunction without the risk of a Type I or Type II error. No correction was therefore necessary.

Running the tests yielded both a statistically significant and practically significant decrease in gross conversions, and neither a statistically nor practically significant change in net conversions.   The sign test similarly revealed a significant difference in gross conversions, and not a significant difference in the net conversions. The new feature matches the experiment’s expectations of a decrease in gross conversions and no change in net conversions. This suggests that the new feature does set user expectations differently without reducing the payment rate at the statistically significant levels. However, the net conversion results’ confidence interval exceeds the negative level of practical significance, indicating that there is a risk of a decrease of net conversions, and therefore revenue, should this change be implemented. Given this risk, I would not recommend implementing the new feature.

Recommendation

This experiment tested the hypothesis that a change in the user’s experience of the website which filtered the users according to their stated availability resulted in lower rate of early cancellation. There was a negative result for the experimental group in the effect size test for gross conversions, and a failure to reject the null hypothesis at both the practical and statistical significance levels for net conversions. The new feature resulted in a decrease of gross conversions and a failure to reject the hypothesis that net conversions changed. Further, the confidence interval for the net conversion metric exceeded the negative interval of practical significance. Launching the new feature is risky, as there is a chance that it may decrease revenue. I would therefore recommend not launching the new feature and sticking with the feature in the control settings, and to test other hypotheses, given sufficient time and budget.

Follow-Up Experiment

Framing user expectations is critical to user engagement. I would propose testing the implementation of a prompt at the ‘Start Free Trial” guaranteeing, for instance, a personal tutor who would be assigned to a student upon enrollment.   This employee could offer their help through video check-ins, help with projects, and more generally discussions about programming. This may help to set user expectations when navigating to the ‘Start Free Trial’ page. Implementing this change may be more costly for Udacity, so a commensurate change in the course price could make up the difference, a change which which may require additional testing. This could significantly reduce the number of early cancellations, improve the student’s learning experience, and increase Udacity’s revenue.

The proposed changes to the webpage would have similar invariant metrics and evaluation metrics. The invariant metrics would likely be cookies or user IDs--IDs are more stable once a user has enrolled, and they could track when a person enrolls or completes the course as the unit of diversion. The evaluation metrics would be gross and net conversion, as the former is the proxy for the rate of enrollment in the free trial, and the latter is the rate of those who paid after enrolling for 14 days. Because the follow-up experiment tests the same propositions as its precedent, the null and alternative hypotheses would not change--that is, for the null hypothesis, there is no change in gross and net conversions. The alternative hypothesis would be confirmed if net conversion doesn’t decrease in the experimental group while gross conversion does.

Resources

N/A