DAND Project 7: A/B Test by Vitor Bellini

Introduction

This project is about an A/B Test runned by Udacity. The experiment want to reduce the number of early cancellations on paid courses. To do this, an free trial screener was incorporated on the student enrolling workflow. This screen, after the student click “start free trial”, asked how much time they had available to devote to the course. If the student indicated 5 or more hours per week, they would be taken through the checkout process as usual. If they indicated fewer than 5 hours per week, a message would appear indicating that Udacity courses usually require a greater time commitment for successful completion, and suggesting that the student might like to access the course materials for free.

The hypothesis was that this 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.

Experiment Design

Metric Choice

Invariant Metrics

- **Number of cookies:** Cookies should be randomly assigned to the experiment and control groups since it is the unit of diversion. Therefore it is not expected significant different values between the experiment and control groups.
- **Number of clicks:** The experiment screen appears only after the click on “Start Free Trial”, so this metric should not be affected by the experiment and the values on the experiment and control groups should be equivalents.
- **Click-through-probability:** This metric is the number of clicks divided by the number of cookies. Since both of this metrics are measured before the experiment screen, it is not expected any difference between the two groups, making this metric a good choice for invariant metrics.

Evaluation Metrics

- **Gross conversion:** This metric should be affected by the experiment because the screen after the click on “start free trial” button should affect the user behavior on enrolling the course by setting clearer expectations for students upfront reducing the number of student who left the free trial period. The numerator (enrolling) should decreases on the experiment group while the denominator (clicks) holds constant.
- **Net conversion:** This metric should be affected by the experiment because it is expected that the screen helps to select students with higher chance of past the free trial period and even finish the course. Since it is expected that the number of enrolling decreases, the number of user-ids who past the free trial should not drop significantly.

Not Used

- **Retention:** This metric have different unit of analysis and diversion which could increase the variability. Also, this trend can also be captured by the net conversion metric.
- **Number of user-ids:** This metric have different unit of analysis and diversion which could increase the variability. Also, this information is considered on the gross conversion metric and is not appropriate as invariant since the number of enrollments should be different between groups.

In order to launch the experiment it is expected both the results:

1. Practical significance decrease on gross conversion;
2. No practical significance decrease on net conversion.

Measuring Standard Deviation

Evaluation Metric	Analytical Standard Deviation	Comments
Gross Conversion	0.0202	In this metric I expect the analytical and empirical variances to be similar due to equivalents units of diversion and analysis (cookies).
Net Conversion	0.0156	In this metric I expect the analytical and empirical variances to be similar due to equal units of diversion and analysis (cookies).

Both metrics are probabilities, having the expected distribution of binomial (normal). To this type of distribution, the analytical formula to compute standard deviation is:

$$SD = \sqrt{p * (1 - p) \div N}$$

Since both metrics has the same unit of analysis and unit of diversion, it doesn't seems necessary the empirical data to estimate the variance. This condition lead us to expect equivalent values on both the analytical and empirical variances.

Sizing

Number of Samples vs. Power

The Bonferroni correction will not be used in the analysis phase. Since only two evaluation metrics will be assessed, the impact on false positives would not be that much. Also, the metrics are correlated and the Bonferroni correction would be too conservative for this case and for each metric is defined a practical significance boundary, making the analysis even more rigorous. It seems that the use of the Bonferroni correction would make very difficult to any change to be considered on the experiment.

For $\alpha\ =\ 0.05$ and $\beta = 0.2$ here's the calculation for the number of pageviews needed to power the experiment appropriately:

Gross conversion pageviews needed:

$$\begin{aligned}
&\mathit{minimum\ detectable\ effect}\ (\mathit{dmin})\ = 0.01\\
&\mathit{baseline\ conversion\ rate}\ = 0.20625\\
&\mathit{Number\ of\ clicks}\ =\ 25,835\\
&\mathit{Number\ of\ pageviews\ by\ group}\ =\ \mathit{number\ of\ clicks}\ *\ \mathit{ratio\ pageviews\ by\ clicks}\\
&\qquad\qquad\qquad\qquad\qquad\qquad =\ 25,835\ *\ (3200\ \div 40000)\ =\ 322,937.5\\
&\mathit{Number\ of\ total\ pageviews}\ =\ \mathit{number\ of\ pageviews\ by\ group}\ *\ \mathit{number\ of\ groups}\\
&\qquad\qquad\qquad\qquad\qquad =\ 322,937.5\ *\ 2\ =\ 645,875
\end{aligned}$$

Net conversion pageviews needed:

$$\begin{aligned}
&\mathit{minimum\ detectable\ effect}\ (\mathit{dmin})\ =\ 0.0075\\
&\mathit{baseline\ conversion\ rate}\ =\ 0.1093125\\
&\mathit{Number\ of\ clicks}\ =\ 27,413\\
&\mathit{Number\ of\ pageviews\ by\ group}\ =\ \mathit{number\ of\ clicks}\ *\ \mathit{ratio\ pageviews\ by\ clicks}\\
&\qquad\qquad\qquad\qquad\qquad\qquad =\ 27,413\ *\ (3200\ \div 40000)\ =\ 342,662.5\\
&\mathit{Number\ of\ total\ pageviews}\ =\ \mathit{number\ of\ pageviews\ by\ group}\ *\ \mathit{number\ of\ groups}\\
&\qquad\qquad\qquad\qquad\qquad =\ 322,937.5\ *\ 2\ =\ 685,325
\end{aligned}$$

The higher pageviews estimation, from net conversion, contemplate the power needed on both metrics. So the number of pageviews needed to power the experiment appropriately is **685,325**.

Duration vs. Exposure

To avoid the experiment to take too long, I would divert 100% of the traffic for the experiment. This means that the experiment would run for **18 days**. This decision can only be made due to the interpretation of the low risk of the experiment on the user experience and the Udacity revenues.

Here's the facts that leads to a low risk experiment:

- Even though the user could be identified after enrolling, this information is not necessary for the experiment and should be hidden. So no further sensitive information is collected.
- The experiment only affects the enroll button action. All other Udacity pages would be the same.
- The experiment should affect the number of enrolls, maybe decreasing it. But even if it decreases, it should not interfere or even increase the number of students that pass the trial and make some payments.

Experiment Analysis

Sanity Checks

Invariant Metric	Lower bound	Upper bound	Observed	Passes
Number of cookies	0.4988	0.5012	0.5006	Yes
Number of clicks	0.4959	0.5041	0.5005	Yes
Click-through-probability	-0.0012	0.0013	0	Yes

The sanity check was applied on the three invariant metrics above with 95% of confidence interval. All three observer values were between the confidence interval, showing that the control and experiment groups were properly balanced.

Result Analysis

Effect Size Tests

Evaluation Metric	Lower bound	Upper bound	Statistical significance	Practical significance
Gross conversion	-0.0291	-0.012	Yes	Yes
Net conversion	-0.0116	0.0018	No	No

*Confidence interval of 95% without Bonferroni correction

Sign Tests

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

*Confidence interval of 95% without Bonferroni correction

Summary

For this analysis, the Bonferroni correction was not used. This method would be too conservative and the changes difficult to identify. Also the metrics are correlated and there are only two, leading that the increases of the probability of producing false negatives is not too high. Another factor is that the practical significance increases even more the magnitude of the change to be significant.

The sign tests were compatible with the effect size tests. The p-value for the gross conversion is lower than the chosen alpha of 0.05 and shows that this result to the gross conversion is unlikely to happen by chance. On the other hand, the net conversion p-value was much higher than alpha, meaning that it could happen by chance and this difference could not be attributed to the experiment. This result also is compatible with the net conversion effect size test.

Recommendation

With the result analysis, my recommendation is to not launch the experiment and dig depper on the data. The gross conversion had practical significant decreases, showing that the experiment had actually helped to reduce the number of enrollments, but this was only desirable without a significant drop on the net conversion. The net conversion lower bound (-0.0116) was lower than the practical significance of -0.0075 and the experiment could lead to a revenue drop higher than the accepted by the business. Therefore, the experiment results doesn't match both the launch criteria established by the experiment design.

One option is to test some cohorts and try to identify better results on narrowed groups.

Follow-Up Experiment

A candidate experiment could be the insertion of a live chat support on the logged in page. This could be used by the student to improve understanding of the course materials or questions about the course operation.

The hypothesis is that an enrolled student on the free trial period with this live chat could get more support and motivation to remain enrolled past 14 days boundary. If this hypothesis held true, the Udacity objective of reduce the number of frustrated students who cancel early in the course could be addressed.

Metric Choice

Invariant Metrics

- **Number of user-ids:** Number of users who enroll in the free trial. This step happens before the experiment and should not be affected, making a good invariant metric.

Evaluation Metrics

- **Retention:** Number of user-ids to remain enrolled past the 14-day boundary (and thus make at least one payment) divided by number of user-ids to complete checkout. If the hypothesis hold true, the number of user-ids to remain enrolled past the free trial should increases.
- **Time spent, given enroll:** Time spent logged in divided by number of user-ids to complete checkout. If the student is more motivated and engaged on the course it is

likely that he spent more time logged in and the experiment group presents growth in this metric.

The unit of diversion would be a user-id. Since the experiment will be only on the logged page and objective is to compare the student behavior on the trial period, this unit seems the most appropriate.

Resources

- http://www.evanmiller.org/ab-testing/sample-size.html - calculate sample size
- https://www.graphpad.com/quickcalcs/binomial1.cfm - sign and binomial test