P7: Design an AB/Test

Experiment Design

Metric Choice

Invariant Metrics:

- **Number of cookies**

Number of cookies is our Unit of Diversion. We are going to split them as equally as possible between experiment and control group, so the experiment won't really affect it. It is some levels on before the experiment happens in the funnel.

- **Number of clicks**

The number of unique cookies to click the "Start free trial" button. This one also happens before the trial screener, so it a variable not affected by our experiment and it stays invariant.

Evaluation Metrics:

- **Gross Conversion**

With gross conversion we can check if there is a significant change in number of users that complete the checkout after submitting and viewing the result of the commitment question. This evaluation metric will help us to check if the experiment has a high impact in the number of users that finish the enrolment. If the conversion is reduced too much, the number of students that go to the next level of the funnel would be lower, and therefore could also reduce the number of students to continue past the free trial. This metric has number of cookies, our unit of diversion, as denominator.

- **Net Conversion**

Our hypothesis is clear, we want to see if we reduce the number of frustrated students without significantly reducing the number of students to continue past the free trial. Net conversion actually measures that metric, number of user-ids to remain enrolled past the 14-day boundary. This has to be also one of our evaluation metrics be able to check our hypothesis.

Not Chosen Metrics:

- **Retention**

Retention is similar to net conversion, because both are used to evaluate the number of users-id that remain enrolled past 14-days. The main difference is that retention uses users-id also as denominator, and not our unit of diversion, cookies, so net conversion is much better option and this one can be ignored.

Retention would also need more than 4.7 million page views. That means that a really long experiment time would be needed. For example, even sending 99% of the traffic to experiment, that would mean more than 100 days of experiment, far longer than a few weeks.

- **Click-through-probability**

This metric is closely related with number of cookies and number of clicks. Therefore, we don't really need it as invariant since we are already using the other two.

- **Number of user-ids:**

The number of user-ids is expected to change with the experiment, since the enroll is done after the modified page is shown, so this is not a good invariant metric. It could be used as evaluation metric for the same reason and used to check the first part of the hypothesis, but it is not the best metric as it is not normalized.

What we expect to see in order to launch the experiment:

The users who enroll will decrease (gross conversion) while the users that stay to remain enrolled past the 14-day boundary (net conversion) will not decrease significantly. This means that we are reducing the number of frustrated students.

Measuring Standard Deviation

Gross conversion: 0.0202

Net conversion: 0.0156

As explained before, both use cookies as denominator. Being cookies also the unit of diversion, there won't be much difference between empirical and analytic estimates.

Sizing

Number of Samples vs. Power

The number of pageviews will be 685325. It is the result got from Net Conversion, since it largest than the result from Gross Conversion. Choosing the largest one, we can use the same sample size for both evaluation metrics.

Duration vs. Exposure

The idea is to not run the experiment for more than just a few weeks. "Few" is a very bad way to define it, would be better to have a real number from the decision makers. But taking in account we only know "few", I will choose 3 weeks as few. So, we need to run the experiment less or equal than 21 days. After several tests, that means 82% of the traffic goes to experiment group.

We are going to base our analysis of the experiment risk based on the definition of minimal risk:

"[...]the probability and magnitude of harm that a participant would encounter in normal daily life"

In this case, we are running a online training site, and the experiment itself it is not going to affect the users' normal day life. We are not dealing with financial or health data, because the information we will ask is the number of hours per week they can spend on the training. There won't be any harm to users. There are no checks being done, so they can lie and continue the signup. Even if that is the case, they won't be affected in any way.

So we can say, that the risk is not beyond the minimal risk. After checking the risk, in case the customer considers to send 100% of users to the experiment group, the time needed to run the test would be 17 days.

Experiment Analysis

Sanity Checks

Number of Cookies
- Lower bound: 0.4988
- Upper bound: 0.5012
- Observer: 0.5006
- **Passes: YES**

Number of Clicks
- Lower bound: 0.4959
- Upper bound: 0.5041
- Observer: 0.5004
- **Passes: YES**

Result Analysis

Effect Size Tests

Gross Conversion:
- Lower bound: -0.0291
- Upper bound: -0.0120
- **Statistical significance: YES**
- **Practical significance: YES**

Interval doesn't include 0. Interval boundary of 0,017 doesn't include dmin of 0.01,

Net Conversion:
- Lower bound: -0.0116
- Upper bound: 0.0018
- **Statistical significance: NO**
- **Practical significance: NO**

Interval includes 0. Interval boundary of 0,0098 includes dmin of 0.0075.

Sign Tests

Gross Conversion: 0.0026
Net Conversion: 0.6776

Summary

The confidence level is 95%, so p-value of 0,025. Gross conversion is smaller, so statistically significant. While net conversion is larger, so not significant

When the number of tests increases, so does the likelihood of a type I error, and Bonferroni could be used. In this case I decided to not use it for several reasons:

- The correction controls for false positives at the expense increased false negatives. In our case, ALL metrics must be satisfied to trigger the launch, so false negatives have greatest impact.

Recommendation

We haven't been able to reject the null hypothesis, so the recommendation is to not launch the change.

The gross conversion show that a smaller number of users will complete the checkout so our experiment is actually reducing the number of users enrolling. There are users that wouldn't have enough time per week to spend on the classes, so we recommend them to not enroll.

Net conversion, user-ids to remain enrolled past the 14-day boundary, shows no statistically significant change. At the same time, the confidence interval includes the negative of our practical significance. The experiment could be reducing the number of students who continue past the free trial, not fulfilling the second part of our hypothesis.

Follow-Up Experiment

There are lot of reasons for a person to leave the free trial. The experiment assumes that it is the frustration of not having enough time, but we don't know if that is the most common reason. Also, we don't know why the experiment description assumes is caused by the time available. There could be too many reasons:

- The prerequisites are not well shown so when the student starts find really difficult to progress.
- The prerequisites are not enforced so people ignore them.
- The student is not native in the language used in the course and is having problems to follow it.
- The first lesson is too complex for most of the people, so maybe should be moved forward.
- Personal reasons.
- Economical reasons.
- etc.

Most of the Udacity Nanodegrees require Python knowledge. The student doesn't need to be an expert but at least know the most basic things and the structure of Python code. In Udacity slack channels I usually see people having problems with python, not understanding how to write some particular code or what an error code means. Having basic knowledge of Python is essential to be able to complete the course.

My followup experiment would be this:

For the experiment group, I would add a mandatory 2 hours Python training that shows the basic structure, flow control and debugging. After those 2 hours, the user will be able to download a short cheat-sheet that summarises everything they will probably need, with links to the most usual documentation pages.

Students without prior idea of Python will get the basics in two hours (without having to go with a full-length course), and those with previous experience can get the course finished in much less time.

Our **unit of diversion** would be User-Id. As with previous experiment, after the enrolment the user will get an ID that we can use to send it to control or experiment group.

Since the idea is still to reduce the number of **frustrated students, net conversion** would be the **evaluation metric**. In this particular case, gross conversion is not needed because the experiment take place **after the enrolment.**

In this experiment, the changes are applied after the enrolment, so the number of User-ids is going to be our invariant metric.