A/B Testing final project report
Udacity Free Trial Screener

1. Introduction

The final project deals with the **design and analysis of the results of an A/B test implemented by Udacity**. The experiment consists in **adding a pop-up window**, **when the students clicks on "start Free Trial"** (on the course overview page of a course), where the user is asked to enter the number of hours per week he / she is ready to devote to the course. Pending on this number being above or below 5 hours per week, the student either keep on with the checkout process as usual, or is informed that a higher time commitment is usually needed for successful completion.

As explained in the Final Project Instructions, the hypothesis is that this might "*reduce the number of frustrated students who left the free trial because they did not have enough time, whithout significantly reducing the number of students to continue past the free trial and eventually complete the course*".

The part of the **customer funnel** that interests us was drawn in Figure 1. Some of the metrics that will be used either as invariant or evaluation metric are also shown. Detail on the choice of these metrics is given in section 2.1.

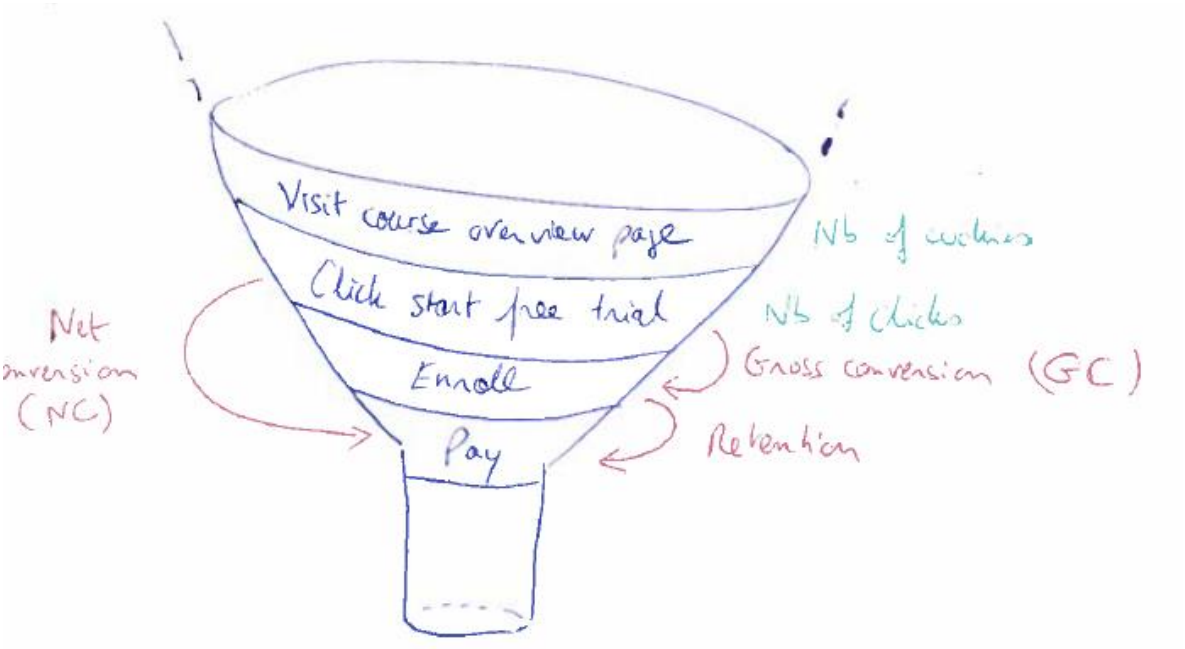

Figure 1: Customer funnel

2. Experiment Design

2.1.Metric choice

The metrics used as invariant metrics are the following:

- **Number of cookies**: number of unique cookies to view the course overview page. This population sizing metrics is well suited for sanity check. Indeed, in the customer funnel the curse overview page view come before the potential click on the "Start free trial" button (which is the starting point for the change that is experimented). This metric shall be comparable in control and experiment groups.

- **Number of clicks**: number of unique coockies to click the "Start free trial" button. Same justification as above: the change is visible to the user only once he/she clicks on the "Start free trial" button, so this metric shall be comparable in control and experiment groups.

The metrics used as evaluation metrics are the following:

- **Gross conversion**: number of user-ids to complete the checkout and enroll in the free trial divided by the number of unique cookies to click the "Start free trial" button. We expect the gross conversion to show a different value in the control and experiment group. Mor eprecisely, the gross conversion is expected to be lower in the experiment group than in the control group. Indeed, it is expected that a fraction of the users clicking on "Start free trial" in the experiment group (those that acknowledge that it will not be possible for them to devote at least 5 hours per week to the course) will switch the the "Access free material" option. The gross conversion seems therefore a meaningful evaluation metrics. Note that because we cannot be sure of the direction of the result, we will use this metric in a two-tailed test rather than one-tailed test.

- **Net conversion**: number of user-ids to remain enrolled past the 14-day boundary (and thus make at least one payment) divided by the number of unique cookies to click the "Start free trial" button. This metric is selected because it allows us to keep an eye on the ratio of users that will make at least one payment, which is key for the financial sustainability of Udacity. Here, we expect the net conversion in the experiment group to be lower than in the control group, but we do not expect this difference to be significant. Again, as we are not sure of the trend, the net conversion metric will be used in a two-tailed test.

NB: The retention (number of user-ids to remain enrolled past the 14-day boundary divided by number of user-ids to complete the checkout) could have also been selected as an evaluation metric. Nevertheless, these three metrics gross conversion (GC), retention (R) and net conversion (NC) are not independent: in fact we have the relation NC = GC * R. For that reason we decided to keep only two of them as evaluation metrics. When estimating the number of samples needed for the experiments (see section 2.3.1), it turned out that using the retention as evaluation metric would require a significantly higher number of samples for reaching the same power.

2.2.Measuring standard deviation

The standard deviation of both the gross conversion (GC) and the net conversion (NC) was **estimated assuming a binomial distribution**:

- s_{GC} = 0.0202
- s_{NC} = 0.0156

Let us consider the gross conversion. This is a ratio whose denominator is expected to be an invariant metric. Its numerator, the number of user-ids to enroll, can be seen as the number of successful outcomes of an experiment consiting of *n* independent repeated trials (*n* being the total number of unique user-ids) that can result in just two possible outcomes: enrollment (success) or no enrollment (failure). By comparing the enrollment of users as a binomial experiment, several assumptions are made: in reality the probability of "success" (that is enrolling) is not the same for each user, and the trials are not independent (for instance if someone that has enrolled is happy with the service provided by Udacity, and then promote it to a friend, this new user would be more likely to enroll). Nevertheless, these assumptions do not seem to be too strong, and they allow us to conveniently have an analytic estimate of the standard deviation for the gross conversion. Besides, this estimate can be compared with the variability empirically assessed either using available data (retrospective analysis) or gathering more data before launching the experiment, if time allows.
The same reasoning apply to the net conversion.

2.3.Sizing

2.3.1. Number of Samples vs. Power

With the chosen evaluation metrics (see section 2.1), we expect the effect size and sign tests to show a significant difference between control and experiment groups for the gross conversion, but not for the net conversion. Therefore, there is no need to have a strict control of the false positives, which we could have done using Bonferroni correction (knowing that gross and net conversion are not independent metrics). Further detail is given in section 3.2.3.
Therefore, **it was decided not to use the Bonferroni correction**. The number of pageviews was estimated using $\alpha = 0.05$.

We used the online calculator presented during the lessons (http://www.evanmiller.org/ab-testing/sample-size.html), with a β of 0.2 as specified in the project instruction notes. The results are shown in Table 1 below:

	Baseline conversion rate	Min detectable effet (dmin)		Nb clicks per variation	Nb pageviews per variation	Total nb pageviews
		absolute	relative			
GC	0.20625	0.01	0.048485	25835	322937	645875
NC	0.1093125	0.0075	0.068611	27413	342663	685325

Table 1: Number of samples

In order to power the experiment appropriately, we need to ensure that we have the maximum total number of pageviews across the different evaluation metrics. In our case, the dimensioning metric is the net conversion , for which it is estimated that **685500 pageviews are needed**. The needed number of samples was also calculated for the retention metric, in the case where it would have been selected together with either the gross conversion or the net conversion. The calculation showed that the retention would have been the dimensioning metric with almost 1000000 needed pageviews. This also explains why the gross and net conversion metrics were chosen as evaluation metrics.

2.3.2. Duration vs. Exposure

It is deemed that the tested change is of medium risk for Udacity:

- On the one hand, we are not dealing with a change that has a technical impact on the behaviour of the service (no server or database migration, no different technical solution for watching videos or answering the quizzes…)
- On the other hand, the change may have a non negligible impact on the revenue expected from the site. If our intuition is false, then we would not like to take the risk of exposing too large a fraction of the traffic.

Based on the above, we decided to **divert 50% of the traffic**. Given the baseline number of pageviews per day of 40000 page views, this means that we expect to divert about 10000 pageviews per day to the control group, and the same amount to the experiment group.

Using this baseline of 40000 pageviews per day, and diverting 50% of the traffic for the A/B test, then **35 days (5 full weeks) would be needed to run the experiment**. Note that fact of using data from an integer number of weeks allows mitigating the potential week days / weekend effect in Udacity traffic.

3. Experiment Analysis

3.1.Sanity Checks

For the sanity check, we use the whole available data (37 days). For the number of cookies and number of clicks metrics, we expect the ratio between control and experiment groups to be close to 0.5. To verify this, we calculate the 95% confidence interval around this value of 0.5 and then check whether the observed ratio is within this confidence interval. The results are shown in Table 2 :

Metric	95% CI lower bound	95% CI upper bound	Observed	Pass sanity check
Nb cookies	0.4988	0.5012	0.5006	yes
Nb of clicks	0.4959	0.5041	0.5005	yes

Table 2: Sanity checks results

3.2.Result Analysis

3.2.1. Effect Size Tests

The results of the effect size tests are shown in Table 3. The confidence interval was computed using data from the 23 consecutive days where enrollment and payment information are available. Indeed, as explained in the project instructions, due to the 14 days period between enrollment and payment, the payment information is not available for the last 14 days of the experiment.

Evaluation metric	95% CI lower bound	95% CI upper bound	Statistical significance	Practical significance
GC	-0.0291	-0.0120	yes	yes
NC	-0.0116	0.0019	no	no

Table 3: Effect size tests results

3.2.2. Sign Tests

The results of the sign tests are shown in Table 4. As for the effect size tests, the sign tests was performed using data from the 23 first days of the experiment.

Evaluation metric	p-value	Statistical significance
GC	0.0026	yes
NC	0.6776	no

Table 4: Sign tests results

3.2.3. Summary

In this experiment **we chose to track 2 metrics**: the gross conversion and the net conversion. We are not expecting the same behaviour from these two metrics: whereas a decrease of the gross conversion is expected in the control group, the net conversion is not expected to show a significant difference between control and experiment groups. If we were to recommend launching the experiment if only one of these two metrics exhibit the expected behaviour, then we would have used the Bonferroni correction as a way to control for the inherent higher number of false positives. This is not the case here: we decided **not to use the Bonferroni correction**, we used an α of 0.05 (i.e 95% confidence interval).

Results of the effect size tests shows that **the proposed change will significantly reduce the gross conversion** (both statistically and practically). **On the other hand, no statistically significant difference is observed in terms of net conversion**. There is **no discrepancies between the effect size hypothesis tests and the sign tests**.

3.3.Recommendation

On the whole the **results** are quite positive in the way that they are **in line with our initial expectations**: implementing the change indeed reduces the gross conversion, but no big effect is seen on the net conversion.

Nevertheless, the evaluation was done using almost half of the advised number of pageviews. Due to the 14 days delay between enrollment and payment information availability the 37 days duration of the experiment corresponds in fact to a duration of 23 days in terms of data, whereas 35 days (6 weeks) were recommended. Increasing the exposure does not seem a valid approach here as we are already diverting 50% of it for a change that is not so transparent for the user and for the business of Udacity.

As a result, I would **recommend not to launch the implementation of the change** until we have at least enough data to get a power of 80% as required.

4. Follow-up Experiment

In order to reduce early cancellations, I would **go further in the idea that the student must know what he/she can achieve given his/her time commitment**. For instance, I would try the following experiment. When the student clicks "Start free trial", there would be a pop-up asking for the hours per week commitment, much like the change we tested in this project. A page will then be displayed with:

- An estimate of the time needed to complete the course of interest (given the time commitment)
- A **list of the courses that could be completed within one month given the entered number of hours per week**.

The hypothesis is that this change would make potential **short terms achievement much more tangible for the students**. This change would also allow the student to **discover other courses** whose content / length / complexity may suit him/her better. As the result, **we expect the number of cancellation with the 14 days following the enrollment to reduce**.

The strategy for the **unit of diversion** would be the **same as used in this project**. Tracking would be done by a cookie until the user possibly enrolls in th efree trial, in which case from that pointforward the user-id would be used.

Because the proposed change allow the user to discover courses that may better suit to its time commitment, the number of pageviews on a given course overview page is expected to change and can no longer be used as an invariant metric. The number of unique cookies to click the "Start free trial" button should still be used for the sanity check. As for **evaluation metrics**, we would use the **gross conversion and net conversion**. The gross conversion would be expected to decrease (as for the experiment we analyzed in this project, and for the reasons given in section 2.1), but this time the net conversion would be expected to increase due to the reduced number of early cancellations.