A/B Test for Anonymous Company

Rodrigo Arriaza, Johannes Burr, Alexander J Ohrt

November 24, 2021

1 Context and Problem to Be Solved

A controlled experiment (A/B test) was conducted to investigate a possible improvement in a mobile game. As players progress through the game, they encounter gates that force them to wait a little while before progressing further. The developers want to implement a new feature - the first gate will be moved from level 30 to level 40. We will investigate if adding this feature would be beneficial.

2 Objectives

Does the proposed new feature lead to an improvement? Three criteria are identified to check if the new feature is an improvement or not: the retention rate after one and seven days respectively and the number of games played during the first 14 days after installment. An A/B test is carried out on each of the measures individually. In addition, a weighted sum of the three measures is constructed as another possible Overall Evaluation Criterion (OEC). The parameter being tested is the level in which a gate is put in the game. As already noted, this experiment introduces two variants, where the first variant is putting the gate in level 30 and the second variant is putting the gate in level 40.

Additionally, an application was built for the management, with the purpose of simplifying testing and interpretation of the test results. The OEC can be chosen, either as one of the three criteria individually or as the weighted sum of the three measures, with weights of choice. Thus, the managers can use their domain knowledge to carry out their desired A/B test and hence make the best possible decision. Perhaps the retention rate 1 day after installing the game is the most important to improve? Or is the retention rate after 7 days more important? Or the number of game rounds played during the first 14 days after installing the app? Or even more complex, perhaps one wants the feature to lead to an improvement in a combination of these three criteria? All these questions can be answered using the application.

3 Data presentation

The data set that this business case is based on is publicly available via Kaggle[1]. The data set contains data for almost 90200 users of the app. The different information elements that were recorded per user are the number of games played during the first 14 days after install and the user retention after 1 and 7 days respectively. Moreover, an indicator of whether the user was in the control group (gate placed at level 30) or in the treatment group (gate placed at level 40) was recorded, which is essential for doing the A/B tests. The users were split almost equally into control and treatment, having 49.56% of the users in the control group and 50.44% in the treatment group. Note that since this is data made by someone else we do not know how the process of assigning variants to the users was done. We have to assume that proper randomization was used, i.e. that each user had an equal chance of being assigned to each variant. We imagine that this was done by randomly assigning each of the users to one of the two groups, where assignment to each group has probability 0.5. This ensures that the control and treatment groups are properly randomized and avoids introducing a bias in the analysis.

[1] https://www.kaggle.com/yufengsui/mobile-games-ab-testing

3.1 Concretization of The Experiment

As already noted, the suggested Overall Evaluation Criterion (OEC) is a weighted combination of three different metrics in the following way

$$OEC = w_1 \cdot R7 + w_2 \cdot R1 + w_3 \cdot SGR, \tag{1}$$

where $\sum_i w_i = 1$ are the weights. In this formula, $R7$ represents the retention rate 7 days after installing the game, $R1$ represents the retention rate 1 day after installing the app and SGR represents the number of games each user played during the first 14 days after first installing the game. In order to make SGR comparable to the binary retention variables, it was scaled to $[0, 1]$. Using A/B testing techniques, the OEC for the two variants were compared, in order to investigate if there are statistical differences in control and treatment. We chose the commonly used statistical significance level of 5% for the calculated p-values. If a p-value falls below this level, we will conclude that the gate placement has a statistically significant effect on the OEC.

4 Results

We begin by comparing the values for 1-day retention, 7-day retention and sum of games played for the given data set.

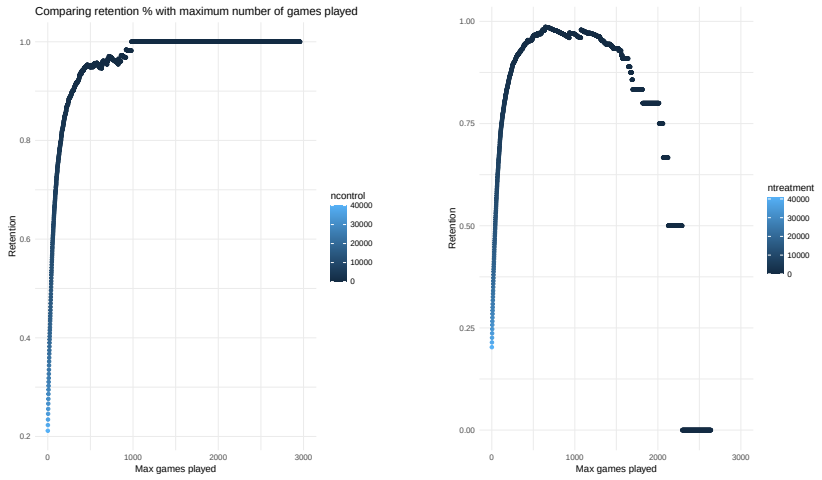
Group	1-day retention	7-day retention	Mean total games
Control	44.82%	19.02%	51.34
Treatment	44.23%	18.20%	51.30
Uplift	-1.32%	-4.30%	-0.20%

Table 1: Comparison of the 3 individual metrics. The percentages represent the percentage of users of the app retaining after 1 and 7 days respectively.

Furthermore we explore the relation between the games played by a given user and the retention rate, as shown in figure 1. As one can see, the more games a user in the control group plays, the higher the retention rate after 7 days is. However, for the treatment group, the retention rate starts to decrease after 1000 games played. Thus, the retention rate after 7 days and the amount of games played the first 14 days look to be correlated, which is expected, where the correlation depends on which of the two groups a user belongs to. But is there a causal relation between the two criteria? We cannot infer this directly from these plots, but it is possible that moving the gate to level 40 makes the game less interesting to players and causes them to dropout. Note that this cannot be concluded from these plots alone and would need further investigation, but it is an interesting question that should be explored.

Furthermore, we proceed to do the statistical tests. Since the retention rates are binary features and the sum of games are skewed towards zero, a normality assumption is not justified. Instead, for testing the OEC a non-parametric Wilcoxon Rank Test is conducted. It does not require a distribution assumption and instead tests the hypothesis whether the Medians of both groups are equal by comparing the ranks after ordering both groups. When the p-value of each statistical test is below the common significance level of 5%, the null hypothesis that the groups do not differ can be rejected.

Here, the individual results are presented. A proportions test did not reveal significant results between the two variants in terms of retention after one day. However, retention after seven days was significantly different between the variants: When the gate was displayed after 40 rounds, 18.2% instead of 19.02% of players kept on playing, which is equivalent to a relative change, or uplift, of roughly -4.3%. Thus, the conclusion when choosing retention after seven days as the OEC is that the new feature represents a deterioration. Also in terms of number of games played, keeping the gate at 30 looks slightly superior in the data. A Median of 17 instead of 16 games were played when being displayed the gate at level 30. With $p = 0.0502$ this result is on the edge of the chosen significance level, so we cannot state with



(a) Control (b) Treatment

Figure 1: Comparison on the 7-day retention rate vs minimum number of games

certainty that the groups differ. However, we can state with 95% confidence that the true shift of median between both groups lies between $[-0.005\%; +0.0017\%]$, which is minor in practice.

5 Conclusions

Based on the data collected, the new feature should not be included. The feature brought no improvement on any individual OEC, yet slightly worsened the results on two of them. Less games were played by users and on average fewer users returned to the game after seven days, when the gate was displayed at a later point in time. The gate should remain at the point 30.

There are, however, some further improvements that could have been done to the experiment. Since we don't know how the data was collected, it may not take into account the Day-of-Week-effect, which could be an important factor when it comes to how much the game is played. Moreover, seasonality is not considered neither. For example, is the game played at different rates during the summer vacation from school compared to during the autumn semester? Would the data look very different depending on which season and which day of the week it is collected? Furthermore, primacy and novelty effects could be considered as well. Was the game completely new at the time of collecting the data? Or was it old and well-known among smartphone users? Such questions cannot be answered with the provided data, but, if one could collect this type of information as well, it could give increased insight to the managers.

6 Code

The code for the analysis and the Shiny App can also be found on GitHub:
https://github.com/arr15334/CookieCatsDashboard