Assignment: Evaluate an AB test of the recommendation algorithm

Introduction

Bambino is a successful e-shop selling toys in two countries: Chille and the New Eldorado. In both countries, Bambino strives to sell as many products as possible, to increase the average size of an order and to maximize the revenue earned.

For this purpose, Bambino has developed a recommendation algorithm that proposes to clients additional products they may like. Once the algorithm was developed, Bambino has decided to run an AB test that would show what additional value the algorithm brings. The AB test was run from 17.5.2023 to 16.6.2023 and was conducted as follows.

Each visitor coming to a Bambino website was assigned a group (an "AB group") on her first visit. The group was either 1 or 2, each assigned with 50% probability (theoretically). The test was set up so that the group was retained in the visitor's next visits. The visitors in group 1 were shown the recommendations, while the visitors in group 2 were not shown any of them. From now on, we shall refer to group 1 as "reco group" and to group 2 as "control group".

Now that the test is over, Bambino wishes to evaluate the results. There are some complicating factors in this.

First, Bambino's IT is not the best on the planet. Sometimes, no AB group is assigned to a web visitor at all. Bambino does not know how often this happens. There is also some suspicion that the probability with which the groups 1 and 2 are assigned is not precisely 50:50.

Second, Bambino tracks its web visitors using the Google analytics (GA) tool. Sometimes — for an unknown reason — an order is not tracked at all. In this case, Bambino receives the order just fine, but it does not know which AB group the order originated from.

Data structure

You have a sample of the relevant Bambino data from the test period. The data consists of two tables:

Table clients_final:

This table tracks all sessions (visits) to the Bambino website during the test period. Each line is one such a visit. The data comes from GA.

- **date** – date of each session in YYYY-MM-DD format. Date spans the period 17.5.2023 to 16.6.2023.
- **country** – Chille (CH) or the New Eldorado (NE)
- **sessionID –** ID of a session (a visit). SessionID per se is not a unique identifier, it can repeat across different users and/or days. It is unique within one day and in a combination with the clientID. One user can have multiple visits, both on the same day and on different days.
- **clientID** – ID of a visitor. You may assume that sessions with the same clientID belong to the same visitor.
- **IsNew** - equals 1 if the session is the first session of a particular visitor. Otherwise it equals 0.
- **abUser** – AB group. Contains values 1, 2 and — in rare circumstances — 99. 99 means that the group was, by a mistake, unassigned.

- **orderNumber** – ID number of a particular order. If no order was concluded within the particular session, the column is empty.

Table orders_final:

This table summarizes the orders Bambino received during the test period. Each line corresponds to one order. The table comes directly from the Bambino's accounting system.

- **country -** Chille (CH) or the New Eldorado (NE)
- **date** - date when the given order was received. Again in a YYYY-MM-DD format.
- **orderNumber** – ID number of a particular order. Corresponds to the same column in the previous table.
- **quantity** – number of products in the order
- **revenue** – total revenue in EUR earned in the order, i.e. the sum of products' prices. It should be positive or zero.

Task

You were asked to evaluate the AB test per each country. In particular, you should answer the following questions:

- Is the ratio of users in the reco group and users in the test group really 50:50? Can you test it by an appropriate statistical test? Do you prefer to test it on a daily basis, or to run one test for the whole period? If you run multiple tests, do you need all of them to have positive results to verify the 50:50 distribution hypothesis?

  No, the ration of users in reco group and is the test group is really not 50:50.
  Here is an explanation:
  - ***I tested the data distribution for each day and also for the whole period.***
    - ***Not a single suitable distribution was found for the country CH, nor within each day.***
    - ***For the NE country, two suitable distributions were found within each day, the remaining days were not suitably distributed and the overall test also came out negative.***
  - ***I used a binomial test.***
  - ***It is not necessary for all partial results to be positive for an overall positive test.***
- What about the users with an unassigned group? Bambino thinks the test is fine if their share is below 0.5%.
  - ***Share of users with an unassigned group for CH:***
    The test is not fine, share is 0.59%
  - ***Share of users with an unassigned group for NE:***
    The test is not fine, share is 0.53%

  I checked this task using ratios and a statistical binomial test. Both methods were negative.

- Do you find any other problems related to a group assignment in the data?
  - ***There is the problem, because in the according to the assignment there should be the number 99, but after checking I find that there is NaN.***

- What about the orders that are not in GA data? What is their share? How do you propose to handle them?
    - ***Share of orders that are not in GA data for CH: 5.52%***
    - ***Share of orders that are not in GA data for NE: 3.36%***
    - ***These values are for my AB test not relevant any more, so thanks "inner join" I actually drop them.***
- Does the "reco group" earn, on average, a greater revenue? Does it have larger orders? Propose appropriate metrics and **visualize** them. Is there any other metric you may wish to evaluate?
    - ***Tests for no country show that individual groups have on average higher revenues or larger orders.***
    - ***I used t-test for statistical testing.***
    - ***You can see the visualization in the attached jupyter file or in the folder "plots" with charts. You can find there:***
        - ***data distribution within groups,***
        - ***outlier detection,***
        - ***histograms of occurrence,***
        - ***visualization of averages***
    - ***Due to the frequency of occurrence, I also visualized the comparison of the two groups for the most frequent number of items in the orders.***
- Do you observe any differences between both countries?
    - ***Answers above or in jupyter file.***
- Optional assignment:
    - Obviously, the users (a their orders) in an unassigned group cast some doubts on the evaluation of the test. If their share is high, the evaluation may be quite unreliable. Can you construct some kind of a confidence-like interval for the above metrics that would show how (un)certain your results are?
    - The expected difference between the reco group and the test group is in the order of magnitude of 1%. If a particularly large order arrives in some group, it can skew the results in favour of this group. Propose a solution to this problem.
        - ***by removing outliers (as I did for the data)***
        - ***z-score***

Method

You may analyze the data by any tool that you know, but we highly recommend Python. We expect you to submit us both the code (with appropriate comments) and a short presentation of the results. We also expect to see some visualizations in any tool of your preference (Excel, Tableau, PowerBI, Matplotlib, etc.)

We wish you a good luck and hope you enjoy the assignment!!!!