Home Test - Data Scientist

# Introduction

The purpose of this test is to evaluate your skills as a creative data scientist and to allow a common ground for discussion in the interview that will take place once completed.

Pointers:

* There are no deliveries needed for question #1.
* You should prepare to present your solutions to a small audience.
* Feel free to prepare graphs, diagrams and every other type of content that will help you present and convey your solutions.
* You will have about 1 hour to present both answers.

If you find that any of the questions are not strictly defined, you may use any assumption you need, but make sure to explain why you used them. For any other question please reach for Eyal - [eyal.david@taboola.com](mailto:eyal.david@taboola.com), 0543281009.

# Question #1 - Creativity and modeling

Assume we were able to track every page view in the “open” internet (social apps/sites are not included) so that for each page view we can fetch and record the following fields:

* Time
* Location
* URL
* URL Title
* Platform and Device

1. Based only on the information above, suggest up to 5 different ways to identify users who:

1. Are searching for a new job.
2. Are sport enthusiasts.
3. Are about to buy a house.

2. Pick one of the ways for one of the segments you suggested above and develop it into a full model - start to finish. Address, explain and detail the following:

* Tools you used
* Scoring and classification models you used
* Statistical significance

**Notice**: You can assume that the data is always available and ignore challenges related to architecture, performance, online/offline calculations etc.

# 

# 

# 

# 

# Question #2 - Statistical analysis and inference.

We are interested in estimating the potential scale of audiences (a.k.a. segments) for a data campaign:

* A data campaign is a campaign that is targeting a closed audience.
* The audiences are known and marked in the DB in advance.
* The empiric way to measure an audience’s quality is correlated with the audience’s CTR - the probability to click on a campaign ad that targets this audience.

A common practice in Taboola is to first run a ‘run-of-network’ campaign without any audience targeting and afterwards look for audiences that significantly overperforms the rest. This allows us to predict a future data campaign’s CTR when targeting for these audiences.

You are given two .csv files:

1. For every (campaign, day) - # of views and # of clicks for the last 50 days.
2. For every (campaign, audience, day) - # of views and # of clicks of this specific subset of users.

We are looking for those pairs (campaigns + segments) that overperforms and show potential. The total number of campaigns is ~100,000 and the number of segments is ~100,000. The files you are given are a sample, that contains the data for about 500 campaigns and a subset of about 100 segments.

Write a script that gets the 100 best couples (campaign, audience) in terms of predicted CTR uplift.

Bonus: Can you suggest a whole other methodology for spotting potential overperforming audiences?

Good luck and godspeed :)