Perengo Data Science Challenge

# Introduction: Perengo (Part of TMP Worldwide)

[Perengo](http://www.perengo.com) (TMP Worldwide) is an automated workforce recruiting tool. Our system monitors and improves the performance along our customer’s application funnels resulting in cost efficient conversions at scale.

Algorithmic job matching makes sure that jobs are seen by the most relevant talent audience. To achieve the most optimal campaign performance, all job seeker engagements are tracked via Perengo Analytics, which captures rich events on web, both desktop and mobile, through a JavaScript pixel implementation on the customer browser.

The Perengo Analytics platform processes and enriches the data, then stores it in Amazon Redshift.

# Data

Please find attached data file containing performance and state data of campaigns during a particular timeframe.

DS\_challenge\_data.csv - This file contains performance data of campaigns that were ran for several customers on a set of inventory sources (ie source of the click).

|  |  |
| --- | --- |
| **Column Name** | **Description** |
| supercampaign\_id | Campaign id |
| date | Date |
| clicks | Total clicks |
| conversions | Total conversions or applications |
| cost | Total cost |
| category\_id | Category which the campaign is assigned |
| industry | Industry to which the customer belongs |
| customer | Customer |
| publisher | Source where the campaigns are run |
| target\_cpa | “Target cost per application”, cost / conversions |

These datasets provide rich information that we will use to inform and optimize future iterations of our service. Our goal is to mine these data sets for performance related insights.

At a high level, our **objectives** are to

(1) maximize conversions at (2) a fixed goal (staying below or around the target cost per application) and (3) with a few other constraints.

The parameters we can manipulate are:

* publishers
* categories
* cost-per-click (bid we submit to a publisher indicating how much we are willing to pay for a click, and we get charged that price when our campaign gets a click)
* \*target CPA (if there is good reason to alter it)

# Challenge Task - Part 1 (70% - 80% of allotted time)

We understand this is just a sample of the data and it does not fully represent the context of the problem. We ask you to list any assumptions you make while performing each part of the exercise to help us understand your approach.

Part I :

1. What are the most important features associated with this data set (those that significantly impact a response variable such as total conversions or CPA)? How predictive are they of our objectives?

*Hint: You might want to perform some exploratory analyses on customer/publisher level to get some intuition of the dataset and find a good approach to address the distribution of different features.*

1. Say we want to run more campaigns (existing ones are expiring); using insights from the data, how do you determine the combinations of customer x publisher x category to use for campaigns in order to optimize for scale (maximizing conversions) and/or performance (lower than or close to target CPA)?

* What types of machine learning/statistical models or frameworks will best tackle this problem?
* What is the outcome of the model and how do you translate that into changes we can implement for running the campaigns?
* How would you validate this model?

1. Cost-per-click (CPC) is among the parameters that we can adjust for campaigns. What changes would you make to the existing campaigns, if any, to **maximize conversions** for each customer?

Part II:

We have a new customer whose campaign is only being run on one publisher (in the data it is **customer = ‘new\_customer’**). We would like to create more campaigns for this customer to increase scale and performance.

What combinations of publishers and categories to use?

# Challenge Task - Part 2 (20% - 30% of allotted time)

Now that you have recommendations on what combinations of publishers and categories to use, how would you set up an experiment to test whether those combinations result in better campaign performance?

1. Describe your experimental approach – be as specific as possible.

*Hint: What is the outcome to be measured? How does the group assignment work? What is the sample size?*

1. What statistical test will you use?
2. How long to run the experiment?
3. How do you reach a verdict on the test?
4. What are the pros and cons of your approach?

Please provide the code/script in the programming language in either Python or R.