CASE STUDY FOR ALGORITHM ENGINEER DATA SCIENTIST

Hi and welcome to the trivago BI recruiting data challenge,

We're happy to hear you are interested in working for trivago! As a first step in the application, we would like to see how you make use of your data science skills.

Preparation time: Maximum 7 days

Submission:

1. A .csv file containing two columns: hotel_id, predicted_number_of_clicks for all hotel_ids where click data is not available.
2. The documented source code you used for this task, in R or Python.

CASE STUDY

Situation:

Data set from about eight hundred thousand hotels

Background information:

Along with this document we have provided you data in a .csv format, containing information about some eight hundred thousand hotels. The columns should be understood as follows:

- **hotel_id:** a number uniquely identifying each hotel
- **city_id:** describes the city the hotel is located in
- **clicks:** the number of clicks the hotel has received in a specific timeframe
- **stars:** the stars of the hotel
- **distance_to_center:** distance (in meters) of the hotel to the nearest city center
- **avg_price_hotel:** average hotel price for per night
- **rating:** average rating on a scale from 0 (worst)- 100 (best)
- **nmbr_partners_index:** describes how many partner websites show rates for this hotel, compared to the average within the city. For example, 1.1 means 10% more than the average while 0.8 means 20% less than the average
- **avg_rel_saving:** average saving users achieve on this hotel by using trivago, i.e. the relative difference between the cheapest and most expensive deal for the hotel
- **avg_rank:** average position the hotel has in the list before filters are applied

Task:

Your task is to use this data to build a model that predicts the number of clicks a hotel will receive, depending on the given parameters.

Evaluation:

Note that the column “clicks” contains many missing values. You will need to estimate/predict these missing click values. Also keep in mind that the data is not “clean”, all columns may be missing data or contain nonsensical values.

Your predictions will be evaluated by a normalized weighted mean square error:

$$error := \frac{\sum_{i=0}^{n} w_i * (predictedClicks_i - observedClicks_i)^2}{\sum_{i=0}^{n} w_i}$$

where

$$w_i := log(observedClicks_i + 1) + 1.$$

Thus: the more clicks a hotel receives the better the model is supposed to be!