LTR Test Task

As the next step of our recruitment process, we'd like you to complete a simple test task.

GOAL

The goal of this assignment is to predict if an image is relevant or not, given a text query.

CONTEXT

The classical search system takes in a query, generates K candidates from a large database, and refines the K candidates to return a final list of “relevant” results.

In this assignment, you are to work on a text-to-image search system where you are provided with generated candidates for some queries, in which you must provide a solution to classify the irrelevant candidates.

In short, it is a binary classification problem where you should predict if a given candidate is relevant or not to the text query.

Evaluation

You are to submit predictions on an unseen test set.

Your predictions will be evaluated using AUC ROC

(https://scikit-learn.org/stable/modules/generated/sklearn.metrics.roc_auc_score.html)

Hint: your final predictions are not restricted to [0,1] or binary, as long as they are ordered.

DATA

- train.feather (your training data)
- test.feather (unseen test set, where you are to make predictions on)
- submission.csv (sample format for your final prediction submission, you need not follow the order in submission.csv, but please provide an “id:predictions” mappings)

You are provided with training data of 733 queries, each with a different number of candidates. The dataset structure consist of html-related features surrounding an image (but you are not

provided with the raw image, only “src”, which you may or may not use it), along with relevance labels:

1. “id” : unique identifier for an image
2. “query” : text query, which is used to determine the relevance of an image
3. “url_page”: webpage where the image is found
4. “src” : the source image url of the image, this is the url for the image itself
5. “title”: title of the “url_page”
6. “alt”: alternate text for the image
7. “is_relevance”: 1 if image is relevant to the query, 0 otherwise

Descriptions for the other features are left out intentionally, except for the fact that they are surrounding html related attributes of the image found in its webpage.

You are free to use any external data sources or libraries, but please provide reference if they occupy a main chunk of your solution.

You are allowed to use heuristics, non-ML related techniques, or even exploit leakage (if any). However, hand-labeling is not allowed.

WHAT WE EXPECT

A Python notebook or script with the following components:

1. Brief description of your approach. Consider detailing what you think might be important for the task. You may find the following points helpful:
    a. EDA, such as interesting findings / frustrations, top features used/engineered
    b. Any reference that you found interesting
    c. Good ideas that worked / did not work
2. Short summary of your results.
3. “submission.csv” containing your predictions on the test set

You may submit your response through the given link in the original email. Good luck.