Data Scientist Take-Home Challenge

Kueski

The objective of this challenge is to assess you ability to:

- perform data manipulation and data pre-processing
- demonstrate awareness of the computations involved
- propose an adequate approach and algorithm for a recommender system given certain data and restrictions
- deal in an appropriate way with possible memory issues
- assess the performance of a recommender system

Below you will find the instructions for this challenge. Good Luck!

Data

The data employed was obtained from a public Kaggle dataset related to financial products recommendation.
Feel free to have a look at the data description on Kaggle, and please, download from there the file `train_ver2.csv`.

Instructions

0.1 Technology used

The programming language used at Kueski for ML projects is mainly Python. We suggest you work on this challenge using Python, although other languages are not excluded if you think there is a justification for using them.

0.2 Presentation format

You can present your challenge in a Jupyter Notebook. If this is not the best option for you, you can use a different tool, but be sure we have access to your code.

1. Exploratory analysis

You might perform exploratory analysis on this data, but **you are not required to present it to us.** In our evaluation, we will focus mainly on your work on the recommender system.

2. Building a recommender system

In this dataset, you have information about different financial products bought by a given customer, at a given point in time. Below we give you instructions about the recommender system you are required to build using this dataset.

The use case for the model

The financial company will send an email to customers on a weekly basis. In the email there will be a list of ordered recommendations for each customer, and the customer will have to scroll down through the list of recommendations in order to see the subsequent ones. They will be able to click on one of the products they are interested in, and that will lead them to a web page offering that product.

One restriction you have is that each email must contain between 3 and 10 recommendations for each customer. You might decide how many recommendations to show for each customer between this range. Explain your reasoning behind your choice.

The algorithm

We know that, given the dataset available, a convenient approach for this problem is to use a multi-class classifier, where each class corresponds to a different product (as it is done by most participants in that Kaggle competition). *However, the objective of this challenge is to assess your skills building recommender systems, so you are required to build a recommender system, instead of a multi-class classifier*.

Building the model

Please, guide us in a clear way about your reasoning and methodology to build your recommender system: what is the target of your model, what type of recommender system did you choose and why, what technical considerations do you think are important for this model to work in a production environment, how would you approach the cold start problem with this model, how do you deal with data leakage, etc.

Evaluating the performance of the model

Define a metric (or metrics) that you find convenient to assess the performance of the model, and explain the reason for your choice.

Also, please outline in a clear way a methodology to evaluate the performance of your model according to that metric (e.g.: what set/sample did you use to evaluate the performance), and also explain the reason behind that.

3. Conclusions

Add some comments summarizing your work. Also, add comments on how you would iterate over the implementation of this model if you would continue working on it.

We wish you success with this challenge!