Data Science Exercise

Scenario

You work in a team that provides online advertising for a range of clients. Recently, the team has begun to collect rich data on each display advert shown to a customer, and whether that advert was clicked on by the customer.

You have been asked to lead the work on this new type of analysis, by developing a model to predict the probability that a given impression will lead to a click. In production, we would hope to use this model to improve the performance of our campaigns which serve millions of impressions per day in Real-time.

Data sets

You have available to you the following data samples to develop your work. When working with our data, we typically use the Python packages *json*, *datetime* and *pandas*; feel free to use those or any other packages you wish.

campaign_impressions.json

A sample log of ad impressions seen by users for the campaigns in question.

Sample file structure

- uuid : unique identifier for the user who saw the ad impression
- ts : timestamp at which the ad impression was shown to the user
- conv : whether the impression led to an in-store conversion
- resp.oi : unique identifier for the client for whom the advertising campaign was run
- resp.cr : unique identifier for the content of the advert shown to the user
- resp.c : unique identifier for the advertising campaign of the impression shown to the user
- dev.os : unique identifier for the operating system of the user's device
- dev.sid : unique identifier for the source of the data
- dev.app : unique identifier of the app on which the user saw the ad impression

user_segments.csv

This is a sample from the user segments database denoting to which segments the user belongs

Sample file structure

- USER_ID : unique identifier for a user (equivalent to the uuid field in the log of ad impressions)
- ANIMAL : whether the user has been identified as having a pet
- CAR_OWNER : whether the user has been identified as owning a car
- GARDEN : whether the user has been identified as owning a garden
- OFFICE_WORKER: whether the user has been identified as working in an office
- PARENT : whether the user has been identified as a parent

The exercise

Your task is to develop the first draft of a predictive model that will be used in a Real-time environment. Your model will have to return a Boolean to the question whether a person will convert after seeing an impression for an advertising campaign or not.

1. Exploratory Data Analysis and Data Preparation

This task should be taken relatively seriously and is expected to be presented in detail as understanding the data is consequential for later design decisions.

Here are some tips from previous interview solutions:

- Data is a little hard to read, but nothing too crazy
- Data Preparation is rather light in this specific exercise
- Feature generation potential is limited, but we are open to surprises. 😉 Ideas for useful complementary data collection are also welcome.

2. Machine Learning

The goal is to develop a machine learning model to predict if an impression will lead to an in-store conversion.

Predictive performance is not relevant, but we'll be interested in the followings:

- Choice of cost metric
- Choice of algorithm
- Separation of train and test data
- The design of the training pipeline
- Maintenance considerations (e.g. automatic retraining)
- Robust evaluation of predictive performance (again not the performance itself)

3. Design deployment in production

We are expecting a system design supported by code for a Real-time environment. This means that, you should:

1. Explain in theory how the final predictive solution would be used to serve a business idea
2. And also, provide working example code that allows simulation of production inference, where the data is coming one row at a time (like in the .json, but without conversion info)

Tips:

- We understand that this might be the most foreign to most candidates, so any solution that works will be accepted
- We are curious to see the time a single response takes on average
- A possible idea would be to use object-oriented programming (representing data as class instances) instead of the usual batch inference.

You will be evaluated by

- The clarity of your code
- Your train of thoughts
- And the flow of your presentation (interactivity, ability to cope with questions)