

Data scientist - home assignment

In this home assignment, we would like to test your understanding of our raw data files and the ability to choose the accurate models for our needs presented below.

The candidate should explore a data and apply machine learning models on it. You will need to use three given data sources:

1. BI data from sales (15 excel files in Drive each represents a single day).
2. Computer vision (CV) results data (json file).
3. External APIs.

There are 2 tasks in this assignment:

CV data clustering:

Given the CV data (behavioral data from the tuna fish shelf, per brand), the candidate needs to choose and apply a machine learning clustering algorithm on it. After performing the clustering, please explain the results. You can use any clustering technique.

Notes:

1. The data will be given to you as a Json file. Each item in the json represents a buyer. The form of a buyer can be considered as a dictionary with the following key-value pairs:
 - “Buyer_id”: string represents the buyer unique ID.
 - “Actions”: array of actions that a buyer performed in front of a shelf. Each element in the array is a dictionary with the following keys:
 - Name: the name of the action
 - Score: the score the algorithm gives to an action.
 - Second_in_video: the time in the video that the action was performed.
 - x1, y1, x2, y2: the coordinates of the human in the frame.
 - “Creation time”: the date and hour that the entity was created in the database.
 - “Age”: the estimated age of the customer
 - “Gender”: the estimated gender of the customer.
2. You should use the data that you think can solve the problem in the best way.

Prediction of the sales:

Given the BI data, you should use a machine learning algorithm to predict the sales of a product. In order to accurate the prediction, you should add new features from external APIs. This API should add significant data that will help the model to predict better results. We call it external influences variables. You need to choose any external influences API that you think can help you to predict the sales better. We call the sales variable “Pidyon”.

After predicting the results using these variables, try to integrate the computer vision data to the model and predict the sales using this data. The CV data will be in the same format explained in the clustering task.

Note:

1. The data will be given as a CSV file. Please see the attached file.
2. External influences variable can be for example calendar event (holidays).
3. Please add 3 new variables for the external influences variables.
4. You will get 15 csvs for train and test. Split it as you think.
5. Using the CV, add two features that you think can help to predict the pidyon.

In the connection between the CV data & sales data, only the "דגים/בשר משומר" category should be analyzed.

General notes:

1. Use Python to perform the assignment.
2. Use BI and CV data in the save time range.
3. The CV is a data sample - not the same amount of data for both data sources.
4. Using database (SQL/NoSQL) is a bunos (you will choose how to use it).
5. Prepare a summary document of the assignment. It should contain details about the chosen methods, difficulties, pros and cons for each method and the used model, etc.
6. Also attach your code. Add comments for code explanation.
7. For any question, feel free to contact us.

Tal: tal@drill-retail.com

Guy: guy@drill-retail.com

Sivan: sivan@drill-retail.com