Data Science Challenge

The purpose of this challenge is to let you demonstrate the way you think and work. The dataset we are providing contains the orders made by customers in one of our applications.

Here's the description of each column:

- **customer_code:** unique id of a customer;
- **branch_id:** the branch id where this order was made;
- **sales_channel:** the sales channel this order was made;
- **seller_code:** seller that made this order;
- **register_date:** date of the order;
- **total_price:** total price of the order (sum of all items);
- **order_id:** id of this order. A order is formed by a set of items;
- **item_code:** code of the item;
- **quantity:** quantity of items, given by item_code, were bought;
- **item_total_price:** total price of items, i.e., quantity* price;
- **unit_price:** unit price of this item;
- **group_code:** which group this customer belongs;
- **segment_code:** segment this client belongs;
- **is_churn:** True, if we believe the client will not come back. For a given customer_code this value is always the same, it means that "today" (the day you are doing this test) this client is a churn.

Question 1 (10 Points)
List as many use cases for the dataset as possible.

Question 2 (10 Points)
Pick one of the use cases you listed in question 1 and describe how building a statistical model based on the dataset could best be used to improve the business this data comes from.

Question 3 (20 Points)
Implement the model you described in question 2, preferably in Python. The code has to retrieve the data, train and test a statistical model, and report relevant performance criteria. Ideally, we should be able to replicate your analysis from your submitted source-code, so please explicit the versions of the tools and packages you are using.

Question 4 (60 Points)

A. Explain each and every of your design choices, you can use jupyter notebooks. (e.g., preprocessing, model selection, hyper parameters, evaluation criteria). Compare and contrast your choices with alternative methodologies.
B. Describe how you would improve the model in Question 3 if you had more time.

PS: Ideally, we should be able to replicate your analysis from your submitted source-code, so please explicit the versions of the tools and packages you are using.

Challenge based on TOTVS Lab GitHub