Type A: FINAL PROJECT NOT INVOLVING LEADA MODULES

DATASET: SEE fabric_softener.zip

This contains a readme file that explains the relevant files and directions about how to properly clean the data.

Objective: build a predictive model for SKU purchases at the household and weekly level. Use the hold-out set to validate the model as in the Hardie and Fader paper (see the readme again). Ideally you would compare different model based on the in and out-of sample predictive checks.

Include: (a) Visualizations of temporal and household-level variation in SKU choices, (b) evaluations of the importance of pricing and promotions, (c) evaluations of the importance of the different attributes for each SKU (to this aim, you may have to code the attributes with an appropriate set of dummies).

Report: 20 pages at most including text, figures and tables. Moreover, include the code in the appendix.

Grading: it will be based on a combination of 1) clarity of exposition 2) soundness of the approach 3) strength of the results and 4) reproducibility (meaning ability to replicate your approach by running the attached code). I will score your project in each of the area on the range between 0 and 5 for a total of 20 points.

You can work alone or in teams (up to 3 people), but you must disclose the people you worked with and you should upload your final report even if you worked in a team (you must upload the report individually).

Type B: Based on LEADA MODULES

This is strictly individual, since the LEADA modules were constructed in this way.

You can opt for either

Python Webscripting: https://www.teamleada.com/courses/intro-to-web-scraping-in-python

or

R + SQL: https://www.teamleada.com/courses/intro-to-analytics-in-sql

You should receive the email from Leada by the end of the day and each of them will be available for 20$.

Final Project Modality:

it will be based on a combination of 1) clarity of exposition 2) soundness of the approach 3) strength of the results and 4) reproducibility (meaning ability to replicate your approach by running the attached code). I will score your project in each of the area on the range between 0 and 5 for a total of 20 points.

Importantly, for 2) I am expecting you to show me how you implemented statistical methods involving clustering, classification and non-parametric methods. I have checked with LEADA that this can be all implemented while solving the final part of the modules.

Type: C: ANALYZE YOUR OWN DATASET

Here I am open, but it should be of difficult equivalent to type A or B. Thus please check out with me asap. I won't allow any Type C unless you have pre-cleared the topics with me.

Modality is still the same: it will be based on a combination of 1) clarity of exposition 2) soundness of the approach 3) strength of the results and 4) reproducibility (meaning ability to replicate your approach by running the attached code). I will score your project in each of the area on the range between 0 and 5 for a total of 20 points.

Hope this is all clear.

Deadline: NOV 15 at midnight.