Data Scientist take-home challenge

Kueski

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

- perform basic data manipulation and data pre-processing
- demonstrate awareness of the computations involved
- drive inferences based on data aggregation
- train and tune ML models
- asses performance of the ML models
- obtaining clear, useful, and business driven insights from data and models

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

Time limit

You should be able to finish this challenge within 8 hours of work. The challenge may be very demanding and it is part of it to consider a feasible/best approach that can be accomplished within 8 hours of your work.

Data

The data employed was obtained from a public Kaggle competition related to Coupons Purchase.
Feel free to have a look at the data description on Kaggle. For your convenience, you will be handed with the necessary files:

- coupon_detail_train.csv
- coupon_visit_train.csv
- prefecture_locations.csv
- user_list.csv

- coupon_area_train.csv
- coupon_list_train.csv
- translations.json

Please use the files we provided to you instead of downloading them yourself. This will avoid you to deal with encodings since we already encoded them for you using UTF-8. So make sure to read them considering UTF-8 encoding. Also, We have provided the `translations.json` file which will allow you to translate japanese characters. This file consists of a hash of the form:

(location name Japanese characters) → (location name english)

Instructions

0.1 Technology used

Use the language/libraries that you find most appropriate to solve this challenge.

0.2 Translation

You should begin by translating Japanese characters in each of the provided files to english, by means of using the provided `translations.json` as mentioned above.

1. Exploratory analysis

You have several tables with different types of information. Explore the data in a way that you obtain knowledge and insights about the product. Present the summary of these findings.

2. Binary classification

Modeling structure

Create a dataframe where each instance (row) corresponds to a tuple (client, coupon). Each column will correspond to a predictive variable. Then, create a column with the response variable for your model. This response variable is defined as 1 in case the client purchased the corresponding coupon, and 0 otherwise.

Note: You have rich information about previous purchases and browsing behavior, feel free to use all the datasets upon your criterion.

Model implementation
Implement a ML model which predicts your response variable using the predictive features you created. Explain all the process you followed to generate/choose the model.

4. Conclusions

Add some comments summarizing your work. Also, add comments on how would you improved it if further time was given to you.

Notes:

1. You will present this challenge in a video call, so adapt its format in a convenient way for that (jupyter notebook, pdf, etc.).
2. You should share with us the code you created for this challenge. We will take into consideration the quality of it.

🙂 Good luck!