Data Scientist take-home challenge
Kueski

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

- basic data manipulation and processing
- computations and inferences based on data aggregation
- modeling and implementation of ML algorithms

and most importantly:

- obtain *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 in about four hours.
Important: due to the size and complexity of the data, someone could spend entire days working on it. The objective of this challenge is that you develop simple work that can be done in a timespan of four hours. Of course, if you want to spend more time working on it it's OK, but we want to emphasize that we prefer you to present simple, good work, than spending an entire week working on it.

Data

We will use data obtained from a public Kaggle competition related to Coupons Purchase https://www.kaggle.com/c/coupon-purchase-prediction. You can take a look at the data description in that website. 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 *these* files instead of downloading yourself the files from the website - they contain special characters and we don't want you to spend too much time dealing with encoding issues.

These files have been saved using UTF-8 encoding, and you should read them using the same encoding. 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

You can use any language/technology you want for this challenge. In Kueski we use mainly Python, and sometimes R.

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. Definition of the problem

Imagine you work for the Coupons Company, and they ask you for help. They have all this data, but have never used it in any way. They are hoping you can make them use this data in a beneficial way for both the company and the costumers.

Define one problem such that, when tackled, there is a potential gain for the company and the clients. For example: "build a model X that predicts Y", "understand client usage of X product", "define the profiles of clients based on X", "understand X attribute from our clients", etc. Try to argument why you think it is a good idea to try to answer such questions or solve those problems.

The problem you choose must be such that:

1. It involves the usage of a predictive Machine Learning model (even if the final answer you are looking for is not the outcome of that model).
2. The proposed solution can (hypothetically) be implemented by the Coupons Company.

For this problem, what metric can you define so that the performance of the proposed solution could be tracked? How would you decide if the proposed solution is better than not implementing it?

Please, state explicitly the problem you defined, and the answers to the questions in the previous paragraph.

2. Exploratory analysis

In order to have defined your problem in the previous section, you should have spent some time exploring the data before. **Create a report which presents useful insights related to the problem**.

3. Attack your problems

Now it's time for you to **work on the problem you chose**. When you are done, **create also a report which describes how you have approached your problem**:

- What techniques did you use, and why?
- What are the performance metrics for your ML model? Why did you choose them?
- How confident are you of the performance estimations of your model? That is, if they were to be deployed/implemented, would you be confident that your performance estimations will hold?

4. Conclusions

Add some comments summarizing your work and the proposed solutions to your problem

Notes:

1. It is important that in the document you create for this challenge, you address explicitly the questions and requests made here.
2. You should share with us the code you created for this challenge. We will take into consideration the quality of it 🙂

Good luck!