Locaria Interview – Data Science

Data Available

The candidate should find attached a series of CSV tables. The following task regards the cleaning, interpretation, and modelling of this data.

Assignment Objective

Feel free to use any tools or technologies to complete the tasks below:

Data Cleaning

There are no restrictions on *how* you clean or reshape the data, but some suggested steps are:

- Convert all prices to GBP
- Join the metrics available in the 'dimensions' table to the train and test set
- Rebalance, rescale, and otherwise process the dataset so that they're appropriate to the models you're using

Exploratory Analysis

- Summary of the dataset, and any characteristics you find interesting
- Please include a few visualizations to help you explain insights into the set, ensuring that the visualizations you include informed how you performed the task

Prediction

The test set doesn't contain price, and we need you to try to predict the value of this variable from the data provided. You are welcome to use any sort of training process for the model, but it will be assessed using **Mean Absolute Percentage Error**.

Please produce the following:

- One model that is easy to interpret and understand
- A model with the best predictive capability for this problem

You are also asked to send us a file containing your best predictions for the Price of the test set before the interview.

The submission file needs to contain the following columns:

- ID
- predicted_price

Tables

1. train_set.csv

Schema:

Price	Numeric
X2	Numeric
X3	Character
X4	Character
X5	Numeric
X6	Numeric
ID	Numeric
Currency	Character

The target variable of this table is Price.
X2 -> X6 are anonymized features.
Finally, every row will have a unique ID and a non-unique Currency code.

2. test_set.csv

Schema:

Price	Empty
X2	Numeric
X3	Character
X4	Character
X5	Numeric
X6	Numeric
ID	Numeric
Currency	Character

The test has identical schema to the training set, but the values of Price are empty.

3. dimensions.csv

Schema:

ID	Numeric
Dimension	Character
Value	Numeric

This table contains extra information relating to the IDs in the training and test set.
Each ID refers to *up to* 3 extra variables to be used for prediction.

4. exchange_rate.csv

Schema:

rate	Numeric
from_currency	Character
to_currency	Character

exchange_rate.csv contains the exchange rate to convert currencies in the training set into a single currency.

Presentation of Results

You will communicate your findings in a 15-20 minute presentation in the interview. This presentation can take any form you prefer, but PowerPoints/PDFs that draw upon your code usually perform better than Notebooks.

Feel free to be technical, as if you're already part of the team and you're presenting your results to your Data Science colleagues.

You are being assessed on both your capability with technology, and ability to communicate results clearly to an audience. As such, your presentation should detail any techniques used along the way, and why you selected them.

We will be looking for clear explanations of advanced techniques and concepts, so feel free to use visualisations or other methods to help with communication.

You will also be asked to explain, as if speaking to a client, how one of your models works. It's not required that you create additional slides for this purpose, but do remember that you're being evaluated on your ability to present clearly as well as your technical ability.

Other Notes

You're being assessed on your understanding and awareness of Data Science techniques, not the application. If you don't have time to include a technique, such as model tuning, please include an explanation of *how* you would have approached it so we know you understand the theory.

We define Mean Absolute Percentage Error as follows:
Mean(Abs((Actual – Predicted) / Actual))

Most candidates prefer to use R or Python for these tasks.