
**Technical Assessment**

**Data Scientist**

**Materials provided for the assessment:**

- Dataset can be found here: http://archive.ics.uci.edu/ml/datasets/Online+Retail#
- This guidance document for the technical assessment

**Expected outputs:**

- Python code or Python notebook
  - Preferably hosted on repl.it https://repl.it/ (create a login and then send over link to code)
  - Alternatively, Github public repo, link to be sent over
  - Last recourse option: send over Python code or Python notebook by email.
- Where analysis wants to be performed in Excel, Excel spreadsheet sent over by email, however, please be cognisant of volume of the dataset.
- A document, format to be determined by the candidate, answering the below questions.

**Criteria for evaluation:**

- It is expected that the code can be run with no assistance (so ensure requirements to run are described in associated documentation).
- It is also expected that the document hosting the responses to the below questions can also be opened and browsed with no assistance.
- Quality of exploratory data analysis.
- Uncovering of trends in the data.
- Ability to perform time series analysis, unsupervised or supervised machine learning on the dataset, with an explanation of validation methodology where applicable.
- Communication of results and decisions being made in the process.
- Quality of visualisations.

**You have 1 week to submit results back to complete this technical assessment.**

**Practical questions:**

**Question 1:**

At first glance, what would be your comments regarding this dataset and its representativeness of real-world datasets? Please elaborate.

**Question 2:**

Please provide some commentary to the exploratory data analysis to be performed on the dataset. What has the exploratory data analysis allowed you to uncover?

**Question 3:**

Now that you have performed some exploratory analysis of the dataset, would you amend the dataset in any way, and how will you ensure traceability of the modifications or enrichments?

**Question 4:**

What questions would you want to answer with this dataset, and how would you go about setting up the analysis for them?

Please provide an example of at least one analysis on this dataset, with explanation of reason for selecting that one to illustrate (time series analysis, unsupervised or supervised machine learning algorithms) and associated visualisations.

**Question 5 – where applicable:**

What would your approach be to validate the model that you have devised on this dataset? Could you please also elaborate on re-sampling and/or re-training mechanisms?

**Question 6 – optional:**

If you have any advanced skillsets in probabilities and statistics, please demonstrate them on this dataset.

**Question 7:**

Please conclude your analysis with a summary of key findings.

**Question 8:**

What skillsets do you have that you feel have not been highlighted by this technical assessment?

**Theoretical questions:**

**Question 9:**

Where do you see the role of data scientist evolving in 5 years' time and what skills are you building in anticipation?

**Question 10:**

What are some of the fundamental steps you'd want to list in the preparation and wrangling of data before applying machine learning algorithms, and where do you see the biggest value in increasing efficiencies in the pre-processing techniques?

**Question 11 – optional:**

Can you please give an example where false positives and false negatives are not equally balanced in criticality?

**Question 12:**

When defining metrics for validation of your model, what is the approach you take for effective design? What are the pros and cons of e.g. MSE or MAE.

**Question 13:**

What are the differentiating factors in selecting unsupervised machine learning techniques versus supervised machine learning techniques? Please provide some examples of use cases that are best suited to one over the other.

**Question 14 – optional:**

In which cases are random forests better than support vector machines? Are random forests better than decision trees?

**Question 15 – optional:**

Why is dimension reduction important and how does principal component analysis come into play?

**Question 16:**

When performing a linear regression, what are the inherent assumptions that are made on the dataset? What are the drawbacks of a linear model, and when can it not apply?

**Question 17 – optional:**

Please list some sources of model bias and how to manage them. What would you do to avoid and/or manage different types of drift?

**Question 18 – optional:**

Why are outliers or long-tailed distributions important to handle? What are appropriate steps to manage extreme event predictions?

**Question 19:**

What methods would you suggest for handling missing data? Please provide some explanation and assumptions to be tested.

**Question 20:**

How would you explain difference between correlation and causation to a stakeholder or an interested party?

**Question 21:**

What do you see as key criteria to be accounted for when looking at opportunities to move advanced analytics and machine learning algorithms into a production environment?

**Question 22:**

How will you persuade the business that improvements to your algorithm are worth implementing?

**Question 23:**

The data and analytics engineering team at Laing O'Rourke will be built on 3 profiles: data analysts, data scientists and data architects. Please provide your understanding of how those roles fit together and complement one another, highlighting any examples you may have come across at any scale.