News UK Data Technology
Data Science Test - Instructions for Candidates

Background and sample data

This test gives you the opportunity to demonstrate your data analysis and predictive modelling skills.

Two data files are provided, which contain fictitious information about a group of customers and their interactions with a news website:

1. ds_test_demog_out.csv:
This is the output of a SQL query run on a customer database and includes customer ids (uid) and a demographic attribute (attribute_1) recorded for these customers.

2. ds_test_log_out.csv:
This is the output of a SQL query run on a database that contains the event logs for a news website. It includes the customer id (uid) of the customer making the interaction on the website, the date of the interaction, and the name of the website section and the website page that they are interacting with.

We will send you a link to the data for you to download.

Test scenario

You are a data scientist working in Data Technology. Rachel, is the deputy editor for one of our news websites and has come to you and explained that she would like to ***improve the user experience for customers visiting the website by showing more relevant, personalised content*** to visitors, based on who they are. In particular Rachel would like to promote certain types of news stories to visitors based on a particular demographic attribute*.

*This is **attribute_1** in the test data, we don’t reveal what this attribute actually is, but image it could be something like *age, gender, owns a dog, address city* etc that users tell us when they sign up to use the news website.

The challenge for Rachel is that attribute_1 is not very well populated in the customer database, and therefore we can only see it for a subset of visitors who come to the website. **Rachel wants to know if you (as a data scientist in Data Technology) can reliably predict attribute_1 for visitors to the website where it is not known for a visitor. If feasible, she wants to use a predicted value for attribute_1 for personalising content where attribute_1 is not available.**

Using the two data sets provided:

1. Assess the feasibility of predicting the value of **attribute_1** where it is missing.
2. Identify which features (that can be extracted from the data) are most important for predicting **attribute_1**.

Assume that you have two audiences for your findings:

- Rachel and her colleagues in the editorial team (who requested the work).
- Your data science and engineering colleagues (as part of a weekly project review).

Outputs required from you

Suggested time for this exercise is 2-3 hours. Your outputs should include:

1. A few bullet points explaining the findings that you would present to each audience (Rachel and the editorial team, and your data science colleagues) and any analysis outputs (charts, tables etc) - please put all of these into a single document (e.g. Word, Google Doc or PDF) - or if you prefer, include in the Notebook below.
2. Analysis code in the language of your choice (e.g. Python, R etc) - commented where possible - either raw or in a Notebook (e.g. iPython)
3. Data sets generated as part of the analysis (including any interim and final data sets e.g. data tables / data frames)
4. Any observations / thoughts you have on the nature of this problem, and ideas on how useful this would be in practice for a news website.

Please send your code and relevant output in a zip file and send to **laura.pettitt@news.co.uk.** If there are any questions, please also reach out to Laura.

Good luck, and have fun!