praelexis

7 Neutron Avenue, Technopark, Stellenbosch, 7600, South Africa Tel +27 21 200 5817 info@praelexis.com www.praelexis.com

Take-Home Assessment for Data Science Role

You will find the data for the exercise here:
https://www.kaggle.com/crowdflower/twitter-airline-sentiment

The task is to build a simple model of the data using the pymc package (https://www.pymc.io/). Pymc is best installed using anaconda and this guide. Install anaconda, create a new environment, and install pymc version 4 or greater using conda and not pip (Note that pymc was, confusingly, previously called pymc3). If you are unable to install pymc contact us to assist.

Note that this task is different from the task of predicting the sentiment as in the kaggle competition. Here you are free to choose which aspects of the data you want to model.

We are interested in understanding your thought process when approaching the problem. We want to see how you approach preparing and exploring the data, model fitting, validation, testing, evaluation, which metrics you use (and why) and critical thinking in evaluating the results. It is likely that you are unfamiliar with the package or with probabilistic programming. This is on purpose, because it allows us to see how you can quickly develop a **minimal** model using unfamiliar tools. Please do not feel you have to create an unnecessarily complicated solution. Finally we would like you to present your findings.

If you are comfortable using the Jupyter notebook it would be good if you can use that.

Please feel free to contact us and discuss ideas as you progress. We encourage you to reference other people's approaches if you use them by adding links in your code. This allows us to distinguish between your own work and existing work.

We would also like to read any articles or write-ups which you have written.

If you save this to a repository then we can have a look at it and give feedback.

Kind regards,

The Praelexis Recruitment Team