Machine Learning Scientist Test

July 12th 2017

Considerations for the test:

* The total time to work on the test is 2 hours.
* This test can be completed by using different platforms: R, Python and Matlab.
* It’s expected that a candidate can complete the test on at least one platform.
* Zip the files generated and send them to: recruitment@obs-gyn.ox.ac.uk

Test

The dataset represents default of credit card clients. It has 30,000 cases and 24 different attributes. Apply 2-3 classifiers (parametric and non-parametric) to predict the target variable which is ‘default’ based on the rest of the features, comparing their results (using appropriate metrics, e.g., AUC, precision, recall, …) and concluding the ideal solution.

Suggested process:

1. Load the data. Find it in ~/Desktop/Test/Data/codedefault.csv. For details of database refer to ~/Desktop/Test/Data/codedefault.txt.
2. Carry out some exploratory analyses (e.g., how various features and the target variable are distributed) - ideally, with rich visualisations.
3. Train a model to predict the target variable (risk of default).
   1. Employ ~3 different models (ideally, a mix of parametric and non-parametric, e.g., GLM and GBM).
   2. Compare the models’ performances (using a list of appropriate metrics, e.g., AUC, precision, recall, …).
   3. Defend your choice of best model (e.g., what are the strength and weaknesses of each of these models?)
4. What more would you do with this data? Anything to help you devise a better (more accurate / reliable) solution? Please demonstrate.
5. Please report your conclusions (including, figures and tables, if any) in a Word document or a Jupyter notebook. You will be given the opportunity to do voice over your report, in the interview – hence no need to a complete written report.
6. Ensure that all the files that you have generated (scripts, data files, etc.) are removed, and the environment of the software you have used is clean in order to leave it ready for the next candidate.