MSD Data Science Recruitment Exercise

(Internship/Fresh Graduates)

Please read ALL of the following instructions before commencing:

Notes:

1. You have 24 hours from the time of receiving this email to complete this exercise.
2. You are allowed to Google or use whatever resources at your disposal. However do note that copying of answers online without understanding them will do you no good (we will find out!).
3. For the choice of tools we allow Octave/Matlab, Python or R. If you are more comfortable with other programming tools, do feel free to check with us. You are also free to use whatever toolbox / packages / libraries that you wish.
4. As a point of reference, this exercise should not take you more than 3-4 hours.

Deliverables:

1. Programming Script (on page 3)
2. Prediction file in .csv format (on page 3)
3. Report (refer to page 4)

Metric:

We will be using the misclassification rate (confusion matrix) as your metric. More information can be found on page 3. However if you are familiar with concepts like AUC – we welcome that as well!

We wish you all the best! Good Luck!

Background information:

This is a typical classification problem in Data Science. For this exercise, your target variable is **readmitted** in **train.csv**. You have >30 days, < 30 days or the patient is never readmitted to the hospital for diabetic patients. Your job as a data scientist is to predict whether this patient will be readmitted to the hospital within 30 days, more than 30 days or otherwise.

More information is available in IDs_mapping.csv and schema.xlsx to explain certain columns. Question marks (?) indicate that the data is either not available or masked.

There will be no questions entertained with regards to the Dataset. You are welcome to make any assumptions and state them in the report.

Requirements / Assessment:

We will judge you based on the following other than the Metric stated:

1. Taking into account missing & noisy data or potential outliers.
2. Go beyond what is given in the dataset to improve your classification error. (Dropping of columns is allowed)
3. Attempts to prevent over-fitting.
4. Tell us what features/patterns leads to a certain classification based on your model.
5. Choice of model and techniques to improve your model performance.

Programming script:

You should make your code as readable as possible (e.g comments, variable/Object names should make sense).

Your script should be in the following order and commented clearly:

1. Loading all your libraries (R), import modules (Python) or state your toolboxes (Octave/Matlab) if you use them.
2. Go beyond what is given in the dataset to improve your classification error. (Dropping of columns is allowed)
3. Attempts to prevent over-fitting.
4. Modeling approach.
5. Make prediction for the test.csv and save into a csv file. For those familiar with metric beyond confusion matrix such as AUC, we accept probability scores/matrix for a multi-classification problem.

Programming Script:

The programming script you used should be named MSD_<intern/FT>_<Name>.<program specific> where intern/FT indicates an internship potion while FT indicates a full time position. For example if your name is James, using R and applying for an internship position it should be MSD_intern_james.R.

Output file:

The output file(s) should be named MSD_<intern/FT>_<Name>.csv, and the format should follow sample_answer.csv provided. **In other words each row in your output file should correspond to the row in the test data**. If you are using the AUC metric, each column should be the probability of the given class labeled clearly.

Report:

Your report should be as concise, clear, and straight to the point as possible. We will not penalize for grammatical errors or language, as long the point is clear. Feel free to use point form or diagram(s) if it helps to convey your idea. So please spend most of your time in presenting your idea and approach.

Your report should also address the 5 points stated on page (2) under Requirements /Assessment **in order**. In addition, please state any assumptions made upfront. Your report **should not exceed 3 pages** (excluding appendix). Do bear in mind that we are looking for quality and not quantity. We would be more impressed by a succinct report that also manages to address the requirements.

We look forward to your submission!

MSD Data Science
