Callstats.io Data Science
Technical Interview Assignment

The aim of the first question is to evaluate the proficiency of the applicant in terms of fundamental but important concepts of data science. It is expected that the applicant has a technical know-how to deal with unfamiliar data, to apply the right data preprocessing, to report the pros and cons and performance of the selected approach and to have an up-to-date overview on the cutting-edge machine learning techniques. The data set is related to the Callstats product, but the problem is mimicking the questions that Callstats currently is trying to solve.

The purpose of the second question is to examine the applicant's understanding of basic but essential mathematical concepts. No numerical calculations are required.

Question 1

There is a csv file named ***dataset.csv***. It includes the columns of predictor variables named ***features*** and a column of cluster labels called ***labels***.

Apply an advanced ***unsupervised*** clustering technique on the columns of features, explain the reason behind your selected method and describe pros and cons of the applied technique compared to other techniques Handle the missing values. Use the labels column for calculating the accuracy of the model estimations.. Explain the performance of the model. Explain your solutions with visualizations.

Python is the preferred language. R and Matlab are also acceptable. Make a concise report including your code. You should spend a maximum of 60 minutes on this question.

Question 2

The covariance matrix of a predictor data set X and the cross-covariance matrix of the predictors X and responses Y are

$$R_{xx} = \begin{pmatrix} 3 & -0.02 & 0.01 & -2 \\ -0.02 & 5 & 1.5 & 0.02 \\ 0.01 & 1.5 & 2 & 0.01 \\ -2 & 0.02 & 0.01 & 2 \end{pmatrix} \text{ and } R_{xy} = \begin{pmatrix} 5.02 & 0.02 & 0.10 \\ 0.01 & 0.01 & -3.25 \\ -0.02 & -0.04 & -2.50 \\ -4.51 & -0.05 & -0.02 \end{pmatrix}$$

respectively. Explain what happens and what kind of prediction results are to be expected when Multivariate linear regression (MLR) is applied. Explain whether MLR could be the right regression technique here or not. What better method you would consider to map the predictor data to the responses? You should spend a maximum of 30 minutes on this question.

Note : Numerical calculations are not required. Try to explain your interpretation and the reason for your proposed regression method in one to three paragraphs**.**

Assignment evaluation:

- The assignment should take no more than 1.5-2 hours to complete (note that you do not need to submit the assignment righawa, you are welcome to take a few days to work on it - just try to keep the active working time around 2 hours)
- Deliverables can be e.g a Powerpoint, a jupyter notebook with comments, or a Word document. Please include your code as well.
- Assignment evaluation criteria: Good fundamentals, conceptual answers, structured reporting