Extended Technical Interview

For this portion of the interview, we are providing you with three datasets and challenges (see below).

During the interview, we would like you to explain your course of action to someone from our Global Data Science team, for example by showing him/her your workflow in a jupyter-notebook-like application that you will share with our colleague during our virtual call.

If you encounter any major obstacle, please feel free to contact your BI HR representative and he/she will be happy to forward your question to the Global Data Science Team.

Wishing you success!

*** Challenge 1: Breast Cancer ***

The data in train_data.csv is related to cancer diagnoses of different types. Each case includes information on the properties (radius, texture and perimeter) of the three most characteristic cell nuclei. Moreover, the age of the person, the date of the diagnose and treatment start, as well as the cancer type is available. The same information is also present in the file test_data.csv, only the cancer type is missing.

For our meeting please look into the data and perform an exploratory data analysis considering the following:

- What are abnormalities in the data?
- Are there any interesting, perhaps unexpected correlations to be found?
- Create a model for predicting the cancer_type

Select an appropriate model and keep its complexity reasonable (number of used features, etc.)

Please send a submission.csv for the cases in test_data.csv at least 24 hours before the interview that includes the prediction of the cancer type of your model. The cases should be in the same order as in the test_data.csv and should only contain the label of the predicted cancer_type. See the sample_submission.csv for format clarification.

*** Challenge 2: Who flies most? ***

Imagine that you are an employee of an airline company. Based on the data that the airline has collected, the marketing department wants to create an advertising campaign. The target groups are those customers who "fly most" and the marketing unit asks you to figure out what these people have in common and how to design a marketing campaign (in form of an email to each participant of this group).

The data about the bare flights is contained in travels.csv. You are provided with many other data sources: the age of the customers (age.csv), their education (education.csv), their political attitude (political.csv), their miles and more membership status (bronze/silver/gold, mam.csv).

Choose the right data sources to join together and answer the questions provided by the marketing team:

- Who are the people who fly most (what do they have in common)?
- Describe very briefly how the content of the email should look like based on the answer of the previous question.

*** Challenge 3: Postal Packages **(do not spend more than 60 minutes on this task)** ***

You are given a dataset with numeric features HEIGHT, WIDTH and DEPTH of postal packages and a numeric target variable PRICE (in USD) that a postal company earned for sending these packages to the respective destination. Compute a linear regression model for predicting the price given the height, width and depth of the package and report on the performance of the model. Please explain your results in detail and do not just print the R^2 score of a linear regression library (you will be asked to interpret the results, i.e. if you receive an MAE of X USD, is that a good model or not? Should you recommend to put it in production?).

Please do not spend more than 60 minutes on this task.