Data Science Test

Thank you very much for taking the time to take this test. It consists of the following tasks:

Relational database with SQL: here you will have to demonstrate your SQL skills to extract information from relational databases.

Data Analysis & Modelling: in this section you will have to challenge yourself with an unsupervised learning exercise and a supervised modelling problem. You can use R or Python for this task.

Please read the explanation of each exercise carefully and try to do your best. Take particular care in answering the technical questions by trying to organise your code well so that it is easily understandable by someone reading it. You should complete and send the test back to us within three days. Consider investing no more than 3 hours for the completion of the tasks. We are looking forward to reviewing it. Good luck!

1.1 Context

Our company works on a mobile application that allows users to communicate between each other through messages and to have access to musical content. For the profitability of this venture, it is important to retain as many users as possible while reducing the number of users that leave the application uninstalling it from their devices (this is called ’Churn’). Thus, an extremely valuable contribution from the Data Science department would be to be able group the users according to their behaviour in order to take better measures to counter churn.

The folder data contains fictitious data about 5000 profiles during May 2021 for users of this hypothetical application. In particular, you can find the CSV files with the columns indicated here below:

• ***message sent train.csv***
user id: integer, unique user identifier
event timestamp: timestamp indicating the instant at which the user sent a message

• ***message received train.csv***
user id: integer, unique user identifier
event timestamp: timestamp indicating the instant at which the user received a message

• ***music played train.csv***
user id: integer, unique user identifier
event timestamp: timestamp indicating the instant at which the user started playing music
minutes: amount of minutes of music played

• ***uninstall train.csv***
user id: integer, unique user identifier
has uninstalled: integer, 1: user has uninstalled - 0: user has kept the app

Similarly, the analogous files can be found in the folder data/test for a different set of 1000 users, with the exception of uninstall test.csv that only contains the user id column.

1.2 SQL

The product team has requested information about user activity. In particular, they are interested in the users' messaging profiles. A relational database is available to gather and compile the required information. A sample of this database consists of the aforementioned tables message sent train.csv and message received.csv. Considering this setup, please provide SQL queries calculating the following:

1. Percentage of users that have listened to music.

2. Percentage of users that have received more messages than they have sent.

3. For the users that have both received and sent messages determine the average daily ratio between received and sent messages.

4. For the same users as the previous point, determine the average number of messages sent between two subsequently received messages.

Suggestion: load the data to Google Cloud's free 'BigQuery sandbox' and take advantage of their free processing for this task (although any other SQL approach is acceptable).

1.3 Data modeling and analysis with Python/R

For this section you can use R, Python or a combination of both, paired with any strategy and method that you deem appropriate for solving the problem. Please provide the updated has uninstalled test.csv file, all programming scripts (including Jupyter Notebooks) and any comments or explanations to support your work.

1.3.1 Analysis

Analyze the data and write a brief report with your findings and any useful insights that you detect in the data.

1.3.2 Segmentation

Devise a way to segment the users into groups so that the Marketing department can develop targeted actions for specific groups. Add a column to have uninstalled test.csv with the result of the segmentation.

1.3.3 Modelling

Build a model with the data in the data/train folder to predict whether a user will uninstall the app. Afterwards, obtain a prediction for the users in data/test/has uninstalled test.csv by adding a column with the prediction. Please use 1 for users removing the application and 0 for users keeping it. Detail the assumptions you make (if any).