CircleUp Data Science Interview take-home problem

Thank you for your interest in our Data Science position at CircleUp. We've created a short take-home problem for you to tackle. Our primary objectives for giving you this take-home problem are to:

- Assess your capabilities in working with data sets.
- Assess your ability to deal with ambiguous problems
- Assess your ability to build simple machine learning models.

Overview of the data for Question 1

Here is a link to the data you'll be working with:

https://docs.google.com/spreadsheets/d/1q-CSF29xKjpqQEE2Lcw9mPYB1IYkun863r9kH-k1VE8/edit#gid=242500391

user_message

You'll see that it is a csv file with 28,371 rows and 4 columns.

Each row in the table represents a 'content creation event' -- Here are some additional details about each column:

user_id - The id of the user in the system
content_created - The date of the event (always between Jan 1, 2015 – Jan 31, 2015)
content_count – A summary count of the total number of content a user created for the event
total_engagement – A sum total of all the positive customer engagement associated with the event

user

You'll see that it is a csv file with 5,317 rows and 2 columns.

user_id - The id of the user in the system
content_created - A internally defined event that can be represented by an binary outcome

user_features

You'll see that it is a csv file with 5,317 rows and 13 columns.

user_id - The id of the user in the system
var _1 … var_12 - 12 variables that represent features we have been able to derive about a user that can be used to predict the above binary outcome (e.g. content_created)

model_test_file

You'll see that it is a csv file with 59 rows and 13 columns - It follows the exact format as the user_features dataset.

Question 1 - Use the user_message dataset

1. Write a function that calculates the total number of content a user had created over the last year and report the users who have greater than 500 pieces of content created.
2. Define a metric and a corresponding function that determines which are the fastest growing users in terms of positive customer engagement over the last year. Report the top 10 users based on the metric that defines "fastest growing user"

Question 2 - Use the user, user_features & model_test_file datasets

1. Write a function that takes as input the user features and outputs the predicted response variable (e.g. content_created) found in the user dataset
2. Report the predicted response for the users in the model_test_file
3. If you use any visualizations/ metrics to validate the model please include them in the report
4. Please explain the reasoning behind the technique you used to build the model

Submission Instructions:

- Please submit all code that you write to complete the above questions. You should include a detailed explanation / reasoning behind the approach you took as inline comments in the code

- You can complete this exercise in whatever language you wish, but you should choose something that most Engineers working on these types of problems would not find obscure. You should submit the code to us such that it can be executed on our computers (please provide execution instructions if necessary).

- Please do not submit a solution in MATLAB or SAS or a solution that is entirely SQL based.

- Data assumptions: Please let us know about important assumptions you make when performing your analysis.

- Data visualizations: If you used any visualization to understand the data, please send us the visualizations along with code used to generate them.