CTRL / SHIFT

21 Media Circle, #06-01 Infinite Studios
Building
Singapore
6779 4506

Dear Candidate,

Thank you for agreeing to participate in our programming challenge. This challenge consists of three machine learning tasks and one optional spark task. You may wish to attempt the first three tasks in order because there are dependencies between these tasks.

We understand that you need adequate time to complete this test, therefore, you will have a week to complete this exercise from the day that the exercise is emailed to you. Since you have a week we are expecting a high-quality analysis and coding standards.

You should complete the exercise using Python or Spark (preferably pySpark) as programming language and the database of your choice (preferably PostgreSQL). Use an available free cloud service (Ex: www.pythonanywhere.com or www.openshift.com) to host the database and your solution. You need to submit the code via Git, including all the SQL you use to setup the DB.

Your final submission will be evaluated based on the quality of the analysis, quality of the code, and creativity. The final submission must include a writeup which explains your solution and documents all the assumptions made during designing the solution. It also should have a brief explanation of how to execute your solution.

Feel free to criticize the tasks in case you don’t see them correct or you need further information on the requirements.

In case of any technical inquiry, please feel free to drop a line to me, using ideas.thelab@ctrlshift.com.

Task #1: Basic Extract-Transform-Loading (ETL)

4 datasets are provided in this task (camp_data.csv, country_code.csv, currency_code.csv, business_vertical.csv). This task is to test your basic ETL skills. The description of camp_data.csv is shown at Appendix 1 below.

1. Join camp_data.csv with the other 3 files to get country, channel_name, and business_vertical values. The join keys are the label columns in the other 3 csv files.
2. Cleanse the data and handle the missing values.
3. Generate total impressions, total clicks, and total conversions by country, channel_name, and business_vertical separately.
4. Select strategy_id =3718750. Show the cumulative sum of impressions and clicks partitioned by channel_name and region over date in ascending order.
5. Perform other types of EDA and report your findings.
6. Insert the dataset into a database using batch insertion.

Task #2: Feature Engineering and Model Building

This task is to test your understanding of machine learning. You would like to predict the performance (impressions, clicks, conversions) of the strategies for each channel in the next week relative to the last data point available assuming total_spend_cpm is 100. Ex: For strategy_id

3718750 the last data point captured in 2018-09-27. Therefore, you should predict the performance in 2018-10-04.

1. Based on the data provided in task 1, explain what kind of model you will use and why you will choose this model.
2. Engineer your features according to your chosen model. You can do feature construction, feature selection, decomposition, or dimension reduction, etc.
3. Train and test your model. Find out whether it is overfitting or underfitting. If yes, how will you overcome this issue.
4. Explain and Evaluate your model performance.

Task #3: Visualization

This task is to test your visualization skills. You will need to use python Matplotlib library, or any other visualization tool to report your findings in task 1 and task 2. Assume you are presenting your results to non-technical people.

1. Use your creativity to visualize the data in task 1 step 3 and step 4.
2. Use your creativity to visualize the prediction you made in task 2.
3. Use your creativity to visualize the other findings you may have discovered from the data.

Task #4: Spark Movie Recommendation (optional)

This task it to test your basic Spark skills.

1. Go to https://grouplens.org/datasets/ to download some data. GroupLens Research has collected and made available rating data sets from the MovieLens web site.
2. Click on MovieLens.
3. Go to recommended for education and development section
4. Download the Full dataset which includes 26 million movie ratings.
5. Use Spark/SparkSQL to sort movies by their popularity.
6. Use Spark/SparkSQL to find the 10 least popular movies of all time.
7. Use Spark/SparkSQL to find the 10 most popular genres of all time.

Appendix 1:

The filed description of camp_data.csv:
date: Date that the performance captured
business_vertical: The business vertical ID
country: Country ID
region: Region ID
strategy_id: The ID of the campaign
channel_name: Channel ID
goal_type: Type of the goal set for the campaign
total_spend_cpm: Amount spend on this campaign on this day
impressions: Number of impressions served
clicks: Number of clicks
conversions: Number of customers converted