**Candidate Name:**

**Date:**

**General overview:**

Please select two of the three challenges provided below and return the results to us in the next three days.

**Additional notes for the candidate:**

* Please document your work and provide us the results along with commented code.
* The candidate’s work will not be used for commercial purposes outside of the interview process.
* We are interested in the both the thought process and approach
* Bonus points for well communicated and presented results
* Use any format you like including PowerPoint, Word doc, excel files, or Jupyter notebooks etc.

**Challenge 1 – Price optimization:**

A pricing experiment have been done by company ABC with the main aim of increasing revenue. In the experiment, the users were grouped into two groups A (66% of the user base) and B (33% of user base), where in group A users were offered with cheaper price while group B users received a higher price.

Business questions based on supplied data:

* What is your recommendation to the company in terms of setting an optimum price for their product?

It’s not working well, there is a low relation between discounts and purchase. The company should be looking for opportunities for discounts looking for the user’s city, source and even the date, hour that the advertisement is shown.

|  |  |  |  |
| --- | --- | --- | --- |
| converted | price | number of records | % of converted purchases |
| 0 | 39 | 198,641 | 98.01% |
| 59 | 112,351 | 98.44% |
| 1 | 39 | 4,031 | 1.99% |
| 59 | 1,777 | 1.56% |

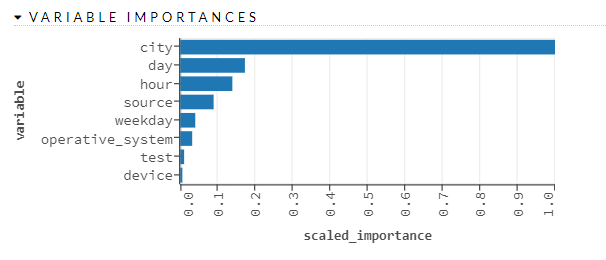
* What are your main findings looking at the data? What is your overall view into user behavior, especially focusing on actionable insights that might increase conversion rate?

Some days in a month at peak time (between 10 to 12 am)

Friend referral has the better purchase rate, on the other hand, direct traffic and ads-bing have the poorest purchase rate.

Mac and IOS followed by windows and android

Cities like New York, Chicago



* Can you optimize the number of days that the test is run? After how many days you would have stopped the test? Why?

The test was ran in 3 months, however the general behavior of users are similar between them. Only running the test for one month is enough to run the test.

Data Details

Table 1: test\_results

**user\_id** : the Id of the user. Can be joined to user\_id in user\_table

**timestamp** : the date and time when the user hit for the first time company XYZ webpage. It is in user local time

**source** : marketing channel that led to the user coming to the site. It can be:

* ads-["google", "facebook", "bing", "yahoo", "other"]. That is, user coming from google ads, yahoo ads, etc.
* seo - ["google", "facebook", "bing", "yahoo", "other"]. That is, user coming from google search, yahoo, facebook, etc.
* friend\_referral : user coming from a referral link of another user
* direct\_traffic: user coming by directly typing the address of the site on the browser

**device** : Can be mobile or web

**operative\_system** : user operative system.

**test**: (i.e. 1 -> higher price) and (0 -> old lower price)

**price** : the price the user sees.

**converted** : 1 -> bought the product - 0 -> left the site without buying it.

Table 2: user\_table

**Columns:**

**user\_id** : the Id of the user. Can be joined to user\_id in test\_results table

**city** : the city where the user is located. Comes from the user ip address

**country** : in which country the city is located

**lat** : city latitude - should match user city

**long** : city longitude - should match user city

**Challenge 2 Recommendation Engine:**

Build a recommendation engine for company ABC’s video streaming product. Currently, the videos shown on their home page to new users are manually chosen and you need to implement a recommendation engine to increase the conversation rate.

You have been presented with the following business questions:

* Classify each video into the following buckets:
  + "Hot" - means trending up. These videos are candidates to be shown.
  + "Stable and Popular" - video view counts are flat, but very high. These videos are candidates to be shown too.
  + "Everything else" - these videos won't be shown.
* What are the main characteristics of the "hot videos"?

The videos classified as hot, are generally videos of **good quality** (1080-480p). Compared to other classes, it has the highest proportion of **videos in English** (~ 34%), on average **shorter videos**, comparable to the "stable and popular".

In comparison to age (publication day), the hot videos are in average a few days more recent than the others (28 days).

\* “Hot”, “Stable and Popular” are in general short videos with good video quality

\* Videos in English are the most “Hot” videos, and Chinese videos are usually the most "Stable and Popular"

* After having identified the characteristics of the hot videos, how would you use this information from a product standpoint?

I believe that with the information of the most important characteristics of video types the company could automate the homepage videos selection for new users. Highlighting the videos that have similar characteristics with "HOT" and "stable and popular", such as short videos with good video quality, videos in English (classified as hot) or Chinese (stable and popular). This approach could also help in selecting and ranking new videos that are published. However we must always be careful about the impact and bias that this type of approach may entail.

Data Details

Table 1: user\_table

**Columns:**

**video\_id** : unique video id

**count** : total count of views for each video per day

**date** : on which day that video was watched

Table 2: video\_features

**Columns:**

**video\_id** : video id, unique by video and joinable to the video id in the other table

**video\_length** : length of the video in seconds

**video\_language** : language of the video, as selected by the user when she uploaded the video

**video\_upload\_date** : when the video was uploaded

**video\_quality** : quality of the video. It can be [ 240p, 360p, 480p, 720p, 1080p]

**Challenge 3: Understanding user and product interaction**

VP of product in the company ABC has asked you to review how users interact with their online travel website.

They store their data in JSON files. Each row in these files lists all the different cities that have been searched for by a user within the same session (as well as some other info about the user).

Business questions:

* There was a bug in the code and one country didn't get logged. Can you guess which country was that? How?

Canada, it’s a lot of cities from Canada in the empty country field and Canada doesn’t appear in the list of countries

* For each city, find the most likely city to be also searched for within the same session.
* Travel sites are browsed by two kinds of users. Users who are actually planning a trip and users who just dream about a vacation. The first ones have obviously a much higher purchasing intent. Users planning a trip often search for cities close to each other, while users who search for cities far away from each other are often just dreaming about a vacation. Based on this idea, can you come up with an algorithm that clusters sessions into two groups: high intent and low intent.

Using latitude/longitude, for each one of the cities researched by user, calculate the pairwise the distance between them and after that extract some aggregated measure from distances, such as the mean, the standard deviation, the minimum and the maximum. Using a clustering algorithm, such as K-mean, we could segregate and visualize different users and combine them to the best high intend and low intend group.

Data Details

Table 1: city\_search

**Columns:**

* **session\_id** : session id.
* **unix\_timestamp** : unix timestamp of session start time
* **cities** : the unique cities which were searched within the same session
* **user** :
  + **user\_id**: the id of the user
  + **joining\_date**: when the user created the account
  + **country**: where the user is based