DATA SCIENCE CHALLENGE DATA SCIENCE

**Video Ranking**

horizontal line

# 

# **Introduction**

## Welcome to our Data Scientist Task!!

Before you dive in, we advise you to read all sections carefully.

Feel free to ask questions or consult with us at any time - just like you would if you were to work here :)

We hope you will enjoy it!

### **The story**

Company XYZ is an online video streaming company, just like YouTube or Dailymotion.

The company has a website, in which they are publishing videos. On the website, videos can be placed in some spots, which can vary in their exposure. So, there are some spots (e.g. top page) that consider more “hot spots” than the other (e.g. bottom page).

This company is interested in knowing whether a video is "hot", stable, or going down. Understanding this would allow us to optimize the videos promoted on the home-page and, therefore, maximize ad revenues.

### **Data modeling**

**The 2 tables are:**

**video\_count** - provides information about how many times each video was seen by day

**Columns**:

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

● count: total count of views for each video

● date: on which day that video was watched

**video\_features** - characteristics of the video.

**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: the 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]

**Examples**

* Let's check one video: how many times was it seen on a given day?

**head (video\_count, 1)**

|  |  |  |
| --- | --- | --- |
| **Column Name** | **Value** | **Description** |
| video\_id | 2303 | Id of the video |
| count | 22 | It was seen 22 times |
| date | 2015-01-07 | On January 7th |

* Let's now check the characteristics of that video 2303:

**subset(video\_features, video\_id == 2303)**

|  |  |  |
| --- | --- | --- |
| **Column Name** | **Value** | **Description** |
| video\_id | 2303 | It is the video we care about. Same as above |
| video\_length | 1071 | the video lasts almost 18 min (1071 seconds) |
| video\_language | Cn | the video is in Chinese |
| video\_upload\_date | 2014-12-10 | was uploaded on Dec, 10 |
| video\_quality | 1080p | video quality is 1080p, i.e. very high |

### **Your Task**

The Head of Product has identified a major problem for the site - a very high home page drop-off rate. That is, the user comes to the home-page and then leaves the site without taking any action or watching any video.

Since customer acquisition costs are very high, this is a huge problem: the company is spending a lot of money to acquire users who don't generate any revenue.

Currently, the videos shown on the home page to new users are manually chosen. The Head of the Product had this idea of creating a new recommended video section on the home page.

She asked you the following:

1. Classify each video into one of these 3 categories:

○ "Hot"

○ "Stable and Popular"

○ "Everything else"

1. What are the main characteristics of "hot videos"? After having identified the characteristics of the hot videos, how would you use this information from a product standpoint?

*Guidelines, tips, and more*

1. You can analyze the data and explore it
2. This data is gold. We can do many things with it. You should think of and list interesting business use cases we can tackle. The use case is something that generates value.
3. The analysis should be clear and well-explained. Every conclusion should be supported by data and written in a way the Product Manager (the one we’re giving service to) will understand.

Having said that, irrelevant information will **not** be received gladly!

1. Try to tell a story, rather than present an academic report. Again, connected to the last dot, people who are not Data Scientists should also understand what you’re saying.

### 

### **FAQ**

***What kind of output should I return?***

The content of the output should have a classification that assigns each video to a certain class, and a justification why each video was segmented to each class.

***Do I have to include an ML algorithm in my analysis?***

No, you can simply analyze the data and do the segmentation based on your analysis.

**What is the meaning of the class’s names?**

The classes names are not that relevant. We will relate to “hot” videos as videos that the probability user will watch them is high and “everything else” as the videos that the probability user will watch them is low. Medium probabilities will be directed to “Stable”.

***How much time do I have to complete the task?***

The task isn’t time stretched and you should have enough time to think and fulfill your dreams. However, life is short, and usually, it takes around 4-7 days to submit home assignments.

***What kind of questions am I allowed to ask?***

Anything! There aren’t any points for not asking questions! Think that we are working together and you are in isolation because of Corona and working from home.

***What sort of resources can I use?***

You can use any kind of resources, materials, tools, or any other reference you wish.