Data Science Internship – Technical Test

Tech Choices

Programming Language: Python

Python is the language I know the best and I've been using on several project, related to Data Science or not.

Libraries:

- Pandas: it is one of the most used Python Data Science Library and I've been practicing it for a few months on a school project's data set.
- Matplotlib: to draw graphics out of the data helping the analysis, works with Pandas Dataframes.
- Scikit-Learn: I quickly tried implementing Machine Learning to the problem. As I am using it on a school project, I used Scikit-learn K-Means method.

Approach to the exploration

Understand the context:

First, I searched more information about the P2P CDN technology to better understand the data and the stakes of the exploration.
I then manually looked at the data: their types, values or ranges of values.

Clean the data set:

I checked the absence of duplicates.
I looked for missing data: only for the last line is the cdn missing. The other values of the line are in the same range as others and I could guess the missing information has nothing to do with a bias: I decided to erase the line.

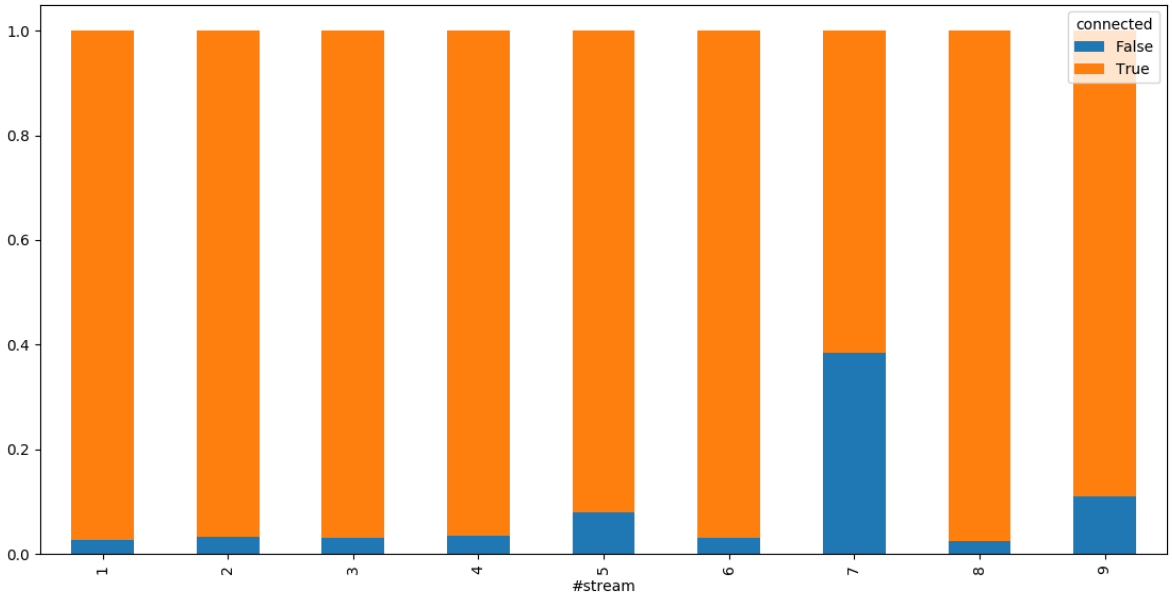
Manipulate the data set:

I first worked using the "connected" boolean to determine in which configurations the misconnections were more frequent. I calculated a connection rate for each type of video, browser and ISP.
I then calculated a ratio of data downloaded through p2p / through CDN and compared it for the different configurations.

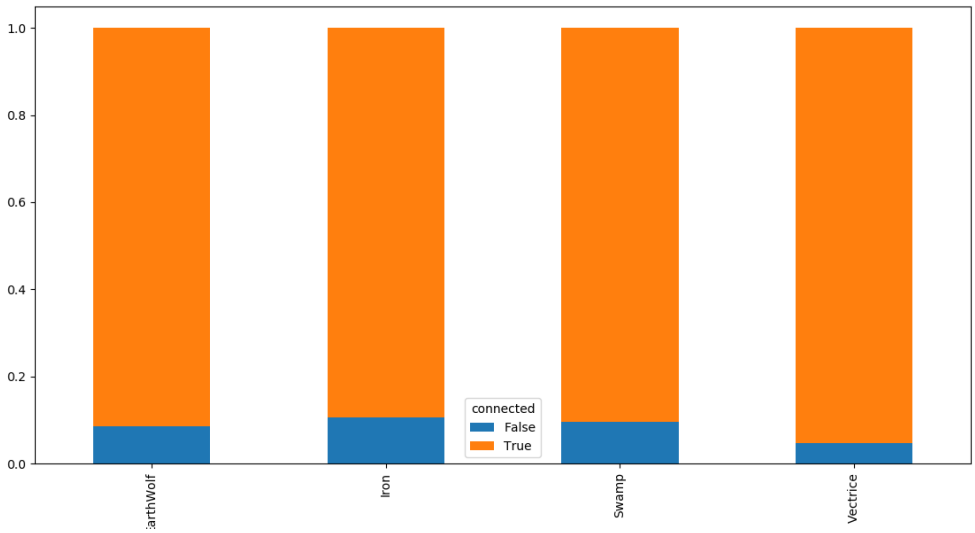
Data-driven recommendation

Connection rate

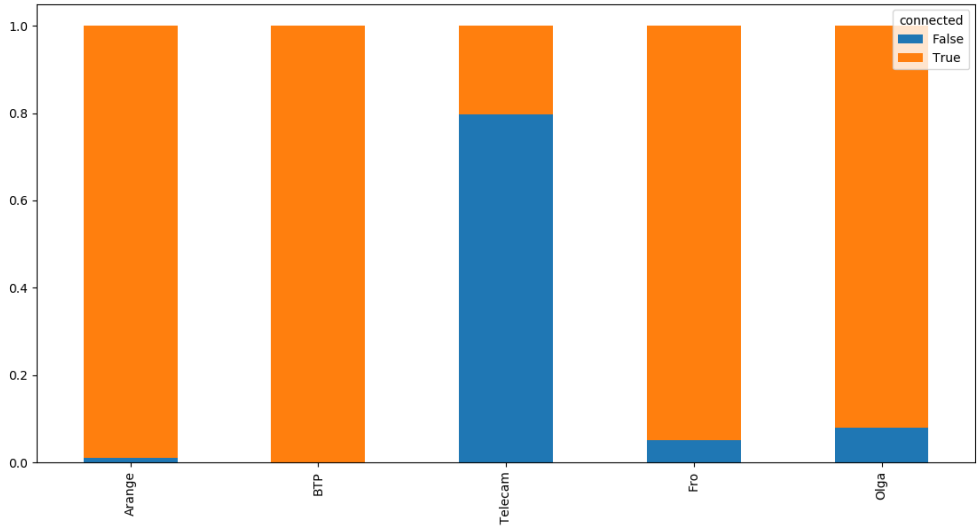
I first compared the connection rate for each video: the following graphical representation of the results shows that the user is more likely to be disconnected from the backend when watching video #7 (about 40% of misconnections while less than 5% on other videos). The service could be improved by finding the reasons of misconnection to this particular video.



Same work with the browser: this time, no significative difference was shown and the rate was quite low for each browser.



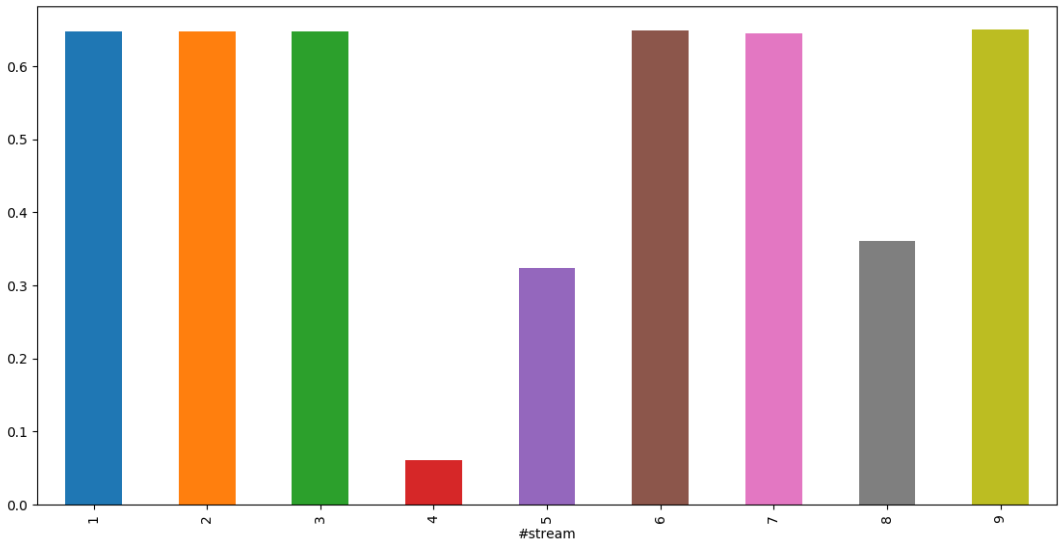
Finally, the same manipulation regarding the different ISP's revealed that 0.8 of users receiving internet from Datch Telecom were disconnected from the service: the product might not be adapted to the company's technical specificities, which could be an improvement.



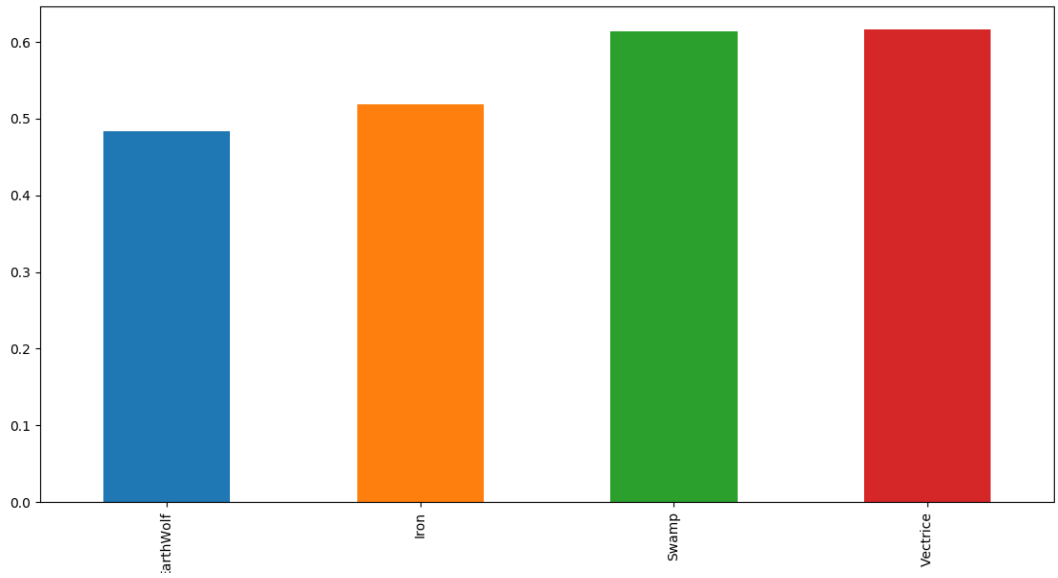
P2P/CDN ratio

I then added to the table a column “ratio” where is calculated the part of data downloaded through p2p (that is p2p / (p2p + cdn)). This only counts the sessions where the user is connected to the backend and the p2p is not null (I wanted to focus on the “connected but 0 p2p cases separately)

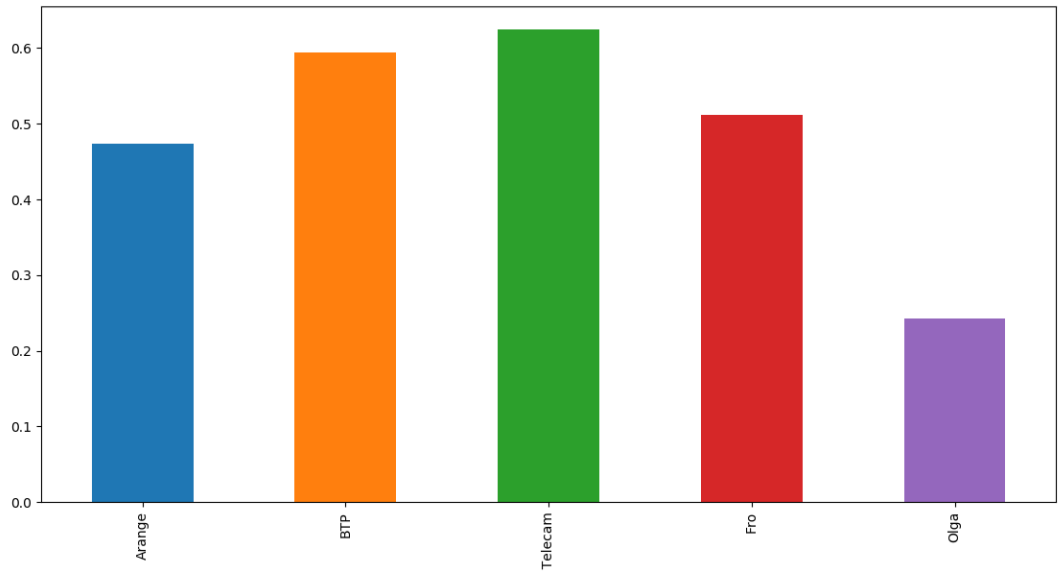
I first compared this ratio for the different videos: this time, video 4 seems to have a very low part of its data being downloaded through p2p. Videos 5 and 8 are also not using the service as much as the other ones.



For the different browsers than, it seems like in the first study that no link is to be found between the efficiency and the browser as their ratios are in the same ranges.



As a last recommendation, I would say to give a further exploration to explain why Olga's users are downloading a much smaller part of their videos through p2p, as shown in the graphic below.



Further work

During the manual exploration, I noticed that in some cases the device was "connected" but still 0 data was downloaded through p2p: I tried to evaluate for which video/ browser/ ISP it happened the most but unfortunately my condition was not working and I couldn't find the mistake yet so this leaves room for improvement.

Also, while working on the connection rates, I briefly tried implementing Machine Learning K-means method. I was trying to find clusters, such as video/ browser/ ISP combinations, showing lower connection. The program was running very slow and not showing much, one of the reasons is probably that the data is not regularized yet. With more time available I would have better prepared the data set to use a ML algorithm.