TFG: Data mining and visualisation to analyse how users consume content in IPTV services

Student: Alex Marin Felices
Director: Miquel Oliver Riera
Tutoring: Alberto Esteve Rivero, Telefónica Chair/Professorship

Introduction

Internet Protocol Television (IPTV) has grown in popularity among users ever since smart TVs and highly sophisticated Internet video streaming services have become more popular. However, as more online content becomes accessible for viewers to pick from, the long-standing problem of choosing what to watch not only persists, but increases. IPTV service providers' recommendations lag behind Over-the-top (OTT) content providers such as Netflix, HBO, Amazon Prime Video, Hulu, etc; which are equipped with powerful recommender systems (RS). OTT providers' content is based on Video on demand (VOD), which allows the user to access specific content, at the time they request it, by viewing it online on their device. IPTV providers, on the other hand, continue to use the Electronic Program Guide (EPG) or Live content. Historically, EPG provides a vast list of channels, but not much RS adoption in IPTV has been observed so far. This can become frustrating and troublesome for users since the list usually contains hundreds of channels with varying types of content [1].

Nowadays we are constantly interacting with RSs. When we search for information, clothes or news on Google, for items on Amazon, movies on Netflix or music on Spotify, they all are using them to rank what we would like the most from their catalogue, or for the thing we are most interested about at that certain moment. These systems take as input data generated by us and other users to output the most accurate recommendations possible [2]. State of-the-art machine learning algorithms have been developed to efficiently create large-scale suggestions to millions of clients, ranging from content-based (CB) RS or collaborative filtering (CF) to more current deep neural networks (DNN). For this study, our focus is going to be put into content available in the IPTV platform from Telefónica, Movistar+. More specifically, in the Latin American provider with contents from Chile.

Objectives of the TFG

1. Validate the initial hypothesis that a high percentage of the consumption made by users in the video service is based on routine consumption by the client and, therefore, it can be identified and anticipated for future recommendations, with the aim of facilitating the content to the user for its consumption. Being able to establish, on a given sample of users, in what percentage this first hypothesis is fulfilled.

2. Identify user consumption patterns in a theoretical model of recommendation that allows us to anticipate said recommendations. The consumption pattern can be

defined upon the type of content seen along the week and for the different timeframes.

3. If the hypothesis of the routine is confirmed given a sample of users, the next step is to clean the data appropriately so that it allows to create an exploratory data analysis (EDA) to:
    - Identify bands of higher utilisation of users.
    - Observe whether user consumption is based on live or Video On Demand (VOD) content
    - Unmask hidden patterns within the data such as: most watched themes/subthemes, prime times, most frequent day of week, etc.
    - Find relationships between subthemes by using Apriori algorithm
    - Create user clusterings based on start time, content theme, subtheme and other segmentation variables.
    - Find differences in content consumption and behaviour based on the type of device.

Motivation

The growing number of streaming services, both live and VOD, are now providing an increasingly diverse range of virtual content. Users are having more difficulty and spending too much time during the selection process as these services and content expand in popularity. Users have negative experiences when deciding what to start watching as a result of this. If you have ever used one of these services, you will probably know how frustrating it can be to search and search for something to watch and eventually run out of time or end up watching something you do not enjoy.

This solution's major purpose is to understand users' behaviour during their watching sessions in order to reduce the amount of time they spend making these decisions. As a result, techniques for uncovering, analysing and interpreting user habits are presented. By having a deeper understanding of these, the aim is to enhance the quality of recommendations both inside and outside the using session, utilising measures such as:

- viewing platform parameters e.g. selection time, viewing sequence (total, partial, serial), device type used (web, PC, smart TV, Internet Protocol Television (IPTV) systems, tablets and smartphones)
- viewing behaviour e.g. start time and end time, frequent content themes and subthemes, number of sessions, average session duration.
- context parameters e.g. individual viewing, as a couple, as a group of friends or as a family; day of the week (workday, weekend), country from where the content is watched

In addition, a possible solution could include comparisons between various recommendation algorithms and diverse ways of displaying them inside the various devices available. In future work, the metrics could then be put to the test using the most prevalent Machine Learning (ML) techniques in order to verify the correctness of these.

Data Results

The dataset consists of around 200.000 sessions from Movistar+ customers placed in Chile between 31.10.2021 and 31.12.2022 (both included). As in any data science project, the majority of the time has been dedicated to cleaning up the data provided, in this case, by Telefónica. This process consisted of different steps. Here you can see some of them:

1. Merging 6 datasets of data, with user sessions coming from 3 different services (ITV, OPL and TVE DTH), and with 2 types (VOD and Live).
2. Change of data types (String to Datetime or Integer)
3. Time conversion from UTC to Chile Continental
4. Dropping of unnecessary, repeated or useless columns
5. Grouping and naming of groups of Themes and Device Types
6. Creation of an auxiliary table to separate all Subthemes in different rows

Once the data has been cleaned I have elaborated some interactive dashboards, using Tableau to filter, analyse and determine the main trends hidden within the data. Here I depict a snapshot from two of the dashboards:

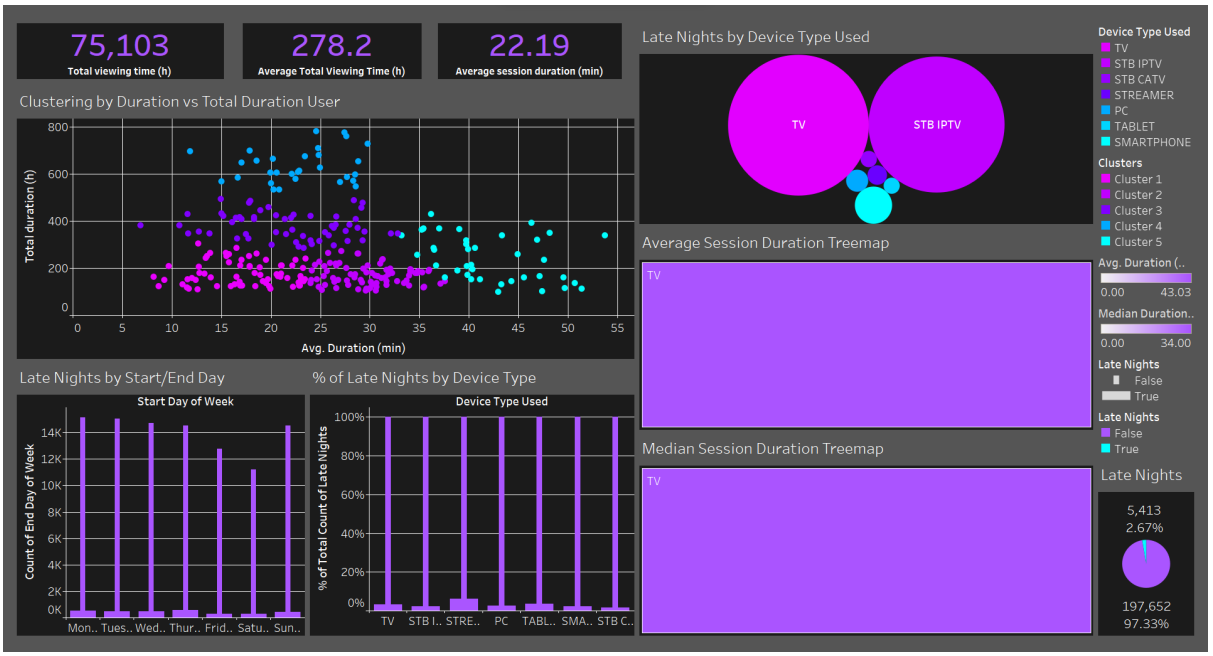

Figure: Dashboard created with the Chile's dataset.

Figure: Dashboard created with the Chile's dataset.

Moreover, I did a clustering analysis to separate the users in different groups based on the starting time of the session, day of the week and the subtheme of the content.

Finally, I have used a common Data Mining algorithm, named Apriori algorithm, to find relationships between subthemes that tend to go with each other. This has also been applied together with starting times, days of the week, device types in different variations.

Bibliography

[1] Li, G., Qiu, L., Yu, C., Cao, H., Liu, Y., & Yang, C. (2020). IPTV channel zapping recommendation with attention mechanism. IEEE Transactions on Multimedia, 23, 538-549.
[2] Milano, S., Taddeo, M., & Floridi, L. (2020). Recommender systems and their ethical challenges. Ai & Society, 35(4), 957-967.