Netflix Movies and TV Shows Clustering

Vidit Ghelani,
Mahesh Lakum
Data Science Trainees,
AlmaBetter, Bangalore

Abstract:

With the advent of technology and the development of computer systems and the internet a vast and varied type of content got accessible to everyone around the world. This also reshaped and redefined the entertainment industry. Netflix which in its initial days was a distributor to movie CDs now with availability of high speed internet, became an online movie and TV shows streaming platform. This exercise is focused upon getting some useful insights of this.

Keywords: EDA (Exploratory Data Analysis), Clustering Technique, Model Training, Machine Learning Algorithms.

1. Problem Statement

This dataset consists of tv shows and movies available on Netflix as of 2019. The dataset is collected from Flixable which is a third-party Netflix search engine.

In 2018, they released an interesting report which shows that the number of TV shows on Netflix has nearly tripled since 2010. The streaming service's number of movies has decreased by more than 2,000 titles since 2010, while its number of TV shows has nearly tripled. It will be interesting to explore what all other insights can be obtained from the same dataset.

Integrating this dataset with other external datasets such as IMDB ratings, rotten tomatoes can also provide many interesting findings.

2. Introduction

A lot has changed in the field of entertainment. Previously a movie or TV show was produced under a specific category with one central idea or theme. Theaters and television were the only media to watch these. The market penetration at that time was quite less due to the limited reach and higher cost. Later on compact storage drives were invented which brought a revolution in this field. With this the market penetration increased and the producers made comparatively high profits. However, here the catch is that piracy had started picking up, limiting the profits of the producers. But nowadays, with the advent of technology in the field of telecommunication, all the above drawbacks have been overcome. Back then each content was charged, but now it's a subscription based model which serves both the viewers and the producers equally well creating a win-win situation for both the parties. With this a wide variety of almost never ending content is available to the users with personalized recommendations to each user based on their fields of interest. This is achieved by applying clustering algorithms to the content. In this exercise we will try to analyze the data and form clusters.

3. Understanding the features

The raw data contains various features mentioned in columns. These features can be divided into demographics, behavioral, and medical history. They are as below:

1. show_id: Unique ID for every Movie / Tv Show
2. type: Identifier - A Movie or TV Show
3. title: Title of the Movie / Tv Show
4. director : Director of the Movie
5. cast: Actors involved in the movie /

show
6. country: Country where the movie / show was produced
7. date_added: Date it was added on Netflix
8. release_year: Actual Release Year of the movie / show
9. rating: TV Rating of the movie / show
10. duration: Total Duration - in minutes or number of seasons
11. listed_in: Genre
12. description: The Summary description

4. Approaching the Problem

The entire process to analyze the data, find out some useful conclusion was done in four steps:

4.1 Pre- Processing
4.2 Performing EDA (Exploratory Data Analysis)
4.3 Performing NLP (Natural Language Processing)
4.4 Model Training
4.5 Conclusion

4.1 Pre- Processing
4.1.0. View the data
4.1.0.1. Understand the Features
4.1.1. Clean the data
We will create a new dataframe and deal with the null values by replacing it with appropriate values or may drop them. This way we do not make any changes to the raw data.

1. Find and drop the irrelevant/ redundant features.
2. Find and deal with the missing values in the dataset.

Summary about the data:

1. show_id : Unique ID for every Movie / Tv Show
2. type : Identifier - A Movie or TV Show
3. title : Title of the Movie / Tv Show
4. director : Director of the Movie
5. cast : Actors involved in the movie / show
6. country : Country where the movie / show was produced
7. date_added : Date it was added on Netflix
8. release_year : Actual Release Year of the movie / show
9. rating : TV Rating of the movie / show
10. duration : Total Duration - in minutes or number of seasons
11. listed_in : Genre
12. description: The Summary description

4.2 Performing EDA (Exploratory Data Analysis)

Here, we plotted various plots for different features. Following are the observations from these plots:

1. Here, we see that of the total content included in this data, the majority of the data corresponds to Movies. TV Shows have about one third the weightage in the data.
2. From the graph to content released per year, we observe that it followed a steady growth till the year 2014. From the year 2015, it just put out an exponential growth.
3. We also observe that in the same year, the production of TV shows also shot up. Uptil then it was almost half that of the number of movies produced. Eventually in the year 2020 we see that TV shows surpassed the total count of movies produced.
4. Looking at the data in the ‘added_year’ column, we see that till the year 2015, quite less amount of movies were added to Netflix. This started rising from the year 2015. Looking at this and the previous plots we can deduce that Netflix did sponsor a lot of these productions. Thus, adding a humongous amount of data.
5. Looking at the month wise distribution of the data added to the platform, we observe that the number of movies and

TV shows added, observed a downward trend from the month of March till July. A reason to speculate could be the fiscal year. The headquarters of Netflix is located in Los Gatos, California, U.S., where the fiscal year is from July to June.

6. Majority of the content is produced in the USA, followed by India, UK and Japan. Also, a lot of the data seems to lack information on the country of origin hence that ambiguity holds true while still not affecting the outcome of our observation.
7. The UK, Japan and South Korea are the only countries where the total number of TV shows produced is more than the number of movies produced.
8. Of all the top content producer nations, India appears to have the least contribution for TV Shows.
9. Of the top directors with the majority of content in the dataset, we observe that “Raul Campos” and “Jan Sulter” top the list with their contribution to movies only. Among these top 15 directors only three directors namely, “Marcus Raboy”, “Quentin Tarantino” and “Ryan Polito” are found to have contributed for both movies and TV shows.
10. TV Shows do not have ratings: 'R', 'PG-13', 'PG', 'NR' and 'G'.
11. The maximum content in the dataset has a rating of 'TV-14' followed by 'TV-PG' & 'R'.
12. 'TV-Y7-FV', 'UR' and 'NC-17' seem to have less or almost no content listed with them.

    Meaning of these ratings:

    TV-14 means the content is unsuitable for children under the age of 14 years.

    TV-PG means the content is unsuitable for younger children.

    R means the content is restricted for viewers under the age of 17 years.

    PG-13 means parental guidance is necessary for children under 13.
13. We see that the proportion of target audience based on the type of content produced by the USA, UK and France is nearly the same. A similar trait was observed for India and Japan. Apparently the cultural beliefs of these two groups are quite closely similar and this is reflected in the type of content produced.
14. The most popular genre is ‘Drama’ followed by ‘Comedy’ and ‘International TV Shows’.
15. Majority of the content is listed under three different categories. The content falling under just one category is about one third of the total data.
16. Top 10 actors based on the number of appearances in movies and TV shows consists of a majority of Indian actors like ‘Anupam Kher’, ‘Shah Rukh Khan’, ‘Naseeruddin Shah’, ‘Om Puri’, ‘Akshay Kumar’, ‘Boman Irani’, ‘Amitabh Bachchan’ and ‘Paresh Raval’.
17. The duration of movies appears to be a near normal distribution with the mean duration of 90 minutes. Looking at the number of seasons for the TV shows, it appears that the majority of the shows have only one season. Hence, the plot is heavily skewed towards the left.
18. These words occur the majority of the time in the titles: ‘Movie’, ‘World’, ‘Man’, ‘Story’, ‘Love’, ‘Christmas’, ‘Day’ and ‘Girl’.

4.3 Performing NLP (Natural Language Processing)

Here, two important features 'description’ and ‘listed_in’ were looked after. We applied NLP to the data of these features. The aim here is to get useful words for better clustering. The task included removing stopwords, punctuations and stemming of words. Thus by doing this we get superior quality data for clustering.

4.4 Training the Models

Here, we feed the data to two different models and then based on the outcome of it we decide the optimum value of the number of clusters.

The models used here are:

- K-means Clustering
  - Silhouette Method
  - Elbow Method
  - Dendogram
- Aglomerative Clustering

5. Challenges Faced:

Few of the challenges we faced were:

1. Need to plot a lot of graphs to analyze the data.
2. Selecting the appropriate number from different clustering plots.
3. When the NLP method is involved, it appears that the process is a bit lengthy.

6. Conclusion:

From this exercise we can conclude that:

1. The available data consists of 69.1% of titles corresponding to movies and the rest 30.9% to TV Shows. Thus, movies as a category of content dominate the quantity in this dataset.
2. Looking at the data from release year 2001 to 2018, the number of movies and TV shows released has observed an exponential rise, with a major break through observed in the year 2014-2015
3. We can also observe that the amount of movies released has been about 2 to 3 times the amount of TV shows released. However, it is evident that this ratio started to reduce from the year 2019. Thus the demand for the TV shows started to peak up from the year 2017 and impacted the production criteria by the year 2019.
4. However, the content started to get on the Netflix platform in mass from the year 2015 - 2016.
5. Looking at the month wise distribution of the data added to the platform, we observe that the number of movies and TV shows added, observed a downward trend from the month of March till July. A reason to speculate could be the fiscal year. The headquarters of Netflix is located in Los Gatos, California, U.S., where the fiscal year is from July to June.
6. Majority of the content producers were from the USA, followed by India, UK and Japan.
7. The UK, Japan and South Korea are the only countries where the total number of TV shows produced is more than the number of movies produced.
8. Of all the top content producer nations, India appears to have the least contribution for TV Shows.
9. Of the top directors with the majority of content in the dataset, we observe that “Raul Campos” and “Jan Sulter” top the list with their contribution to movies only. Among these top 15 directors only three directors namely, “Marcus Raboy”, “Quentin Tarantino” and “Ryan Polito” are found to have contributed for both movies and TV shows.
10. Observation for Ratings:
    TV Shows do not have ratings: 'R', 'PG-13', 'PG', 'NR' and 'G'. The maximum content in the dataset has a rating of 'TV-14' followed by 'TV-PG' & 'R'. 'TV-Y7-FV', 'UR' and 'NC-17' seem to have less or almost no content listed with them.
11. We could say that the majority of the content here is for adults and young adults. Very little content is available for kids.
12. We see that the proportion of target audience based on the type of content produced by the USA, UK and France is nearly the same. A similar trait was observed for India and Japan. Apparently the cultural beliefs of these

two groups are quite closely similar and this is reflected in the type of content produced.

13. The most popular genre is ‘Drama’ followed by ‘Comedy’ and ‘International TV Shows’.
14. Top 10 actors based on the number of appearances in movies and TV shows consists of a majority of Indian actors like ‘Anupam Kher’, ‘Shah Rukh Khan’, ‘Naseeruddin Shah’, ‘Om Puri’, ‘Akshay Kumar’, ‘Boman Irani’, ‘Amitabh Bachchan’ and ‘Paresh Raval’.
15. The duration of movies appears to be a near normal distribution with the mean duration of 90 minutes. Looking at the number of seasons for the TV shows, it appears that the majority of the shows have only one season. Hence, the plot is heavily skewed towards the left.
16. These words occur the majority of the time in the titles: ‘Movie’, ‘World’, ‘Man’, ‘Story’, ‘Love’, ‘Christmas’, ‘Day’ and ‘Girl’.
17. We applied two clustering algorithms namely K- Means and Agglomerative clustering algorithm. The best cluster arrangement we obtained was three.

7. Comprehension:

1. Understanding what type of content is available in different countries?

A: We have plotted a heatmap which shows the type of content produced in each country. This was drawn out from the rating of the movies and TV shows. This has been discussed in detail in point 10, 11 and 12 of the previous section.

2. Is Netflix has increasingly focused on TV rather than movies in recent years?

A: Yes it appears from the graph of release year and added year, that the number of TV shows produced started to increase considerably from the year 2016 and in the year 2020 it was more in quantity than the movies produced.

3. Clustering similar content by matching text-based features?

A: We incorporated a few algorithms and discovered that the optimal number of clusters were three.

References:

1. GeekforGeeks
2. Stackoverflow
3. Python guide
4. Matplotlib guide
5. Seaborn guide
6. Scikit learn guide
7. Github
8. Kaggle