Project Report: Netflix Movies and TV Shows Unsupervised Machine Learning

Project Code link: https://drive.google.com/file/d/1S76YGuet5n-HBrGnlp2Wnx69FwCDNFZu/view?usp=sharing

Presentation link: https://drive.google.com/file/d/14ri0JCT5yBa51nrkvTp4ZSX2J64cuk2V/view?usp=sharing

Professor Name: Khalid Bakhshaliyev

Team Members: (Shiva Karthik Pinjarle Manmohan) and (Ashish Kaleru)

UCID: Sp3254 & Ak3224

1. Introduction

The Netflix Movies and TV Shows Unsupervised Machine Learning project sought to delve into the vast Netflix dataset, leveraging advanced analytics to extract meaningful insights. The primary objectives included understanding content availability trends, exploring patterns through clustering, and implementing a content-based recommender system for enhanced user experience.

Problem Statement

This dataset, sourced from Flixable, encompasses Netflix's catalog of TV shows and movies up to the year 2019. Our objectives include:

1. Exploratory Data Analysis (EDA):
   - Analyzing and extracting insights from the dataset.
   - Identifying patterns, trends, and statistical characteristics.
2. Understanding Content Availability Across Countries:
   - Investigating the distribution of content types in different countries.
   - Gaining insights into the regional preferences for TV shows and movies.
3. Trend Analysis:
   - Assessing whether Netflix has shifted its focus towards producing more TV shows compared to movies in recent years.
   - Examining any discernible trends in the content mix.
4. Clustering Similar Content:
   - Utilizing text-based features to cluster similar content.
   - Applying clustering algorithms to group TV shows and movies with comparable characteristics.

Through these analyses, we aim to unveil key insights into Netflix's content strategy, regional variations, and the evolving landscape of its offerings.

2. Data Exploration

Insights

Chart 1: Alastair Fothergill directed most TV shows in our dataset with a total count of 3 TV Shows. Raul Campos has directed most films in our Movie category with a total movie count of 18.

Chart 2: Takahiro Sakurai acted in most TV shows in our dataset with a total count of 25 TV Shows. Anupam Kher is the most films in our Movie category with a total movie count of 32.

Chart 3: Most TV shows in our dataset are released in only one season almost 1600 and Most films released, have 800 minutes of duration and it is normally distributed.

Chart 4: Documentaries, Stand_up comedy, Dramas, and International Movies is very famous and most content available on Netflix has the same genre.

Chart 5: A notable surge in content production was observed between 2017 and 2020, suggesting a response to global events such as the COVID-19 pandemic.

Chart 6: Geographical concentration was evident, with the United States and India emerging as key contributors to the content pool.

3. Clustering Analysis

The clustering phase aimed to categorize content into distinct groups for a more granular understanding. K-means and DBSCAN algorithms were applied, with the former chosen for its ability to create well-defined clusters. The Elbow method and Silhouette score analysis determined the optimal number of clusters, yielding 12 as the most representative segmentation.

Visualizing the clusters in 2D and 3D plots provided a comprehensive view of the content distribution landscape. This visual exploration allowed for identifying content relationships, patterns, and potential gaps in the existing categorization.

4. Content-Based Recommender System

To enhance user engagement, a content-based recommender system was developed. The system utilized TF-IDF vectorization to convert textual data into a format suitable for analysis. Cosine similarity metrics were employed to gauge the likeness between different shows, enabling the generation of personalized recommendations based on user viewing history.

- TF-IDF and cosine similarity techniques contribute to the accuracy of the recommendations.

5. Conclusion

I have tackled the challenge of unsupervised clustering using the Netflix Dataset. Initially, we employed basic Python code snippets to gain insights into the data, exploring its shape, number of features, data types, statistical information, and more. In the exploratory data analysis (EDA) phase, key findings emerged:

1. Netflix boasts a diverse collection of TV shows and movies, with a notable concentration in the United States.

2. The period between 2017 and 2020 witnessed a peak in Netflix content.

3. The majority of Netflix content originates from the United States, followed by India, which has a higher number of movies and a comparatively lower number of TV shows.

Clustering:

In the clustering phase, we applied the K-means algorithm and explored various cluster numbers. While initially suggesting 12 clusters, further analysis indicated that 9 clusters might be optimal. However, after evaluating the Silhouette score, we concluded that K-Means performed exceptionally well with 12 clusters.

A content-based recommender system was built using the similarity matrix obtained after using cosine similarity. This recommender system will make 10 recommendations to the user based on the type of show they watched.

6. Differentiation from Previous Works

Our project differentiates itself from previous works on platforms like Kaggle through:

- While K-means clustering is commonly used, our project incorporates the DBSCAN and clustering algorithm to explore alternative clustering perspectives, ensuring a more comprehensive understanding of content relationships.

- User-Centric Recommender System:

- The content-based recommender system considers not only content attributes but also user behavior and preferences, providing a more holistic approach to personalized recommendations.

- Incorporation of Insights:

- Our project is concentrating on the content contributors, offering strategic implications for collaboration opportunities and content acquisition strategies.