Netflix Movies And TV Shows Clustering

Project Summary

The entertainment industry is a highly competitive field where success depends on several factors, including genre, rating, production budget, and cast. A recent study was conducted to examine the factors that influence the popularity of movies and TV shows on Netflix. The study utilized a dataset containing approximately 12 variables to cluster movies and TV shows based on audience preferences and popularity. The first step in the analysis involved data wrangling, where missing values were handled, and unique values were checked.

The study found that there were several missing values for the 'director,' 'cast,' 'country,' and 'date_added' columns, which were removed by dropping the corresponding rows. Next, the study performed exploratory data analysis (EDA), where it was discovered that Netflix has more movies (5372) than TV shows (2398) available on its platform. The most common rating for both movies and TV shows is TV-MA, which indicates that Netflix's content caters primarily to adult audiences with a focus on mature and potentially controversial themes.

Furthermore, the study found that the years 2017 and 2018 had the highest number of movie releases, while 2020 had the highest number of TV show releases. The growth rate of movie releases on Netflix was significantly faster than that of TV shows. Although there has been a substantial increase in the number of movies and TV show episodes available on Netflix since 2015, there has been a noticeable drop in the number of movies and TV show episodes produced after 2020.

Based on the countplot, it appears that Netflix adds the highest number of movies and TV shows between October and January, which is the busiest time of year for adding new content to its platform. Netflix has the highest number of contents in the United States, followed by India, where India has the highest number of movies on Netflix.

To cluster the shows, the study focused on six key attributes: director, cast, country, genre, rating, and description. These attributes were transformed into a 10,000-feature TFIDF vectorization, and Principal Component Analysis (PCA) was used to reduce the components to 3000, capturing more than 80% of the variance. Next, two clustering algorithms, K-Means and Agglomerative clustering, were used to group the shows. K-Means determined that the optimal number of clusters was 5, while Agglomerative clustering suggested 7 clusters, which were visualised using a dendrogram.

Finally, a content-based recommender system was created using the similarity matrix obtained through cosine similarity. This system provides personalised recommendations based on the type of show the user has watched, giving them 10 top-notch suggestions to explore.

In summary, the study identified key trends in the Netflix dataset, including the growth rate of movies versus TV shows, the busiest period for adding new content, and the content demographics. Through clustering and a content-based recommender system, the study was able to provide personalised recommendations based on the user's viewing history. The findings of this study offer valuable insights into the factors that influence the popularity of movies and TV shows on Netflix, providing a foundation for further research and analysis in this area.

Contributors Roles:

Vimal Kumar Hoon

1.Data Loading
2.Data handling
3.Handling missing values
4.Data exploration/Visualisation
5.Outliers detection

6.Hypothesis Testing
7.Feature engineering
8.Model deployment

Github Repo link-

https://github.com/vimal-139/Netflix-Movies-and-TV-show-clustering