Netflix Movies And TV Shows Clustering
Project Summary

The entertainment industry is highly competitive, and success is dependent on various factors, including genre, rating, production budget, cast, and more. In this context, a study was conducted to understand the factors influencing the popularity of movies and TV shows on Netflix. The study used a dataset containing around 12 variables to cluster the movies and TV shows based on their popularity and audience preferences. The first step in the analysis involved data wrangling, where missing values were handled, and unique values were checked. The study identified that there were 2389 missing values for the 'director' column, 718 for the 'cast' column, 507 for the 'country' column, and 10 for the 'date_added' column. These missing values were removed by dropping the corresponding rows.

Next, the study performed exploratory data analysis (EDA). The number of movies on Netflix is greater than the number of TV shows, with 5372 movies and 2398 TV shows currently available on the platform. The most common rating for TV shows is TV-MA, indicating that a significant portion of the TV shows available on Netflix are intended for adult audiences. Additionally, TV-MA is the most common rating for both movies and TV shows, suggesting that Netflix's content caters to a primarily adult demographic, with a focus on mature and potentially controversial themes.
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 is significantly faster than that of TV shows. Since 2015, there has been a substantial increase in the number of movies and TV show episodes available on Netflix. However, there has been a notable drop in the number of movies and TV show episodes produced after 2020. It appears that Netflix has given more attention to increasing its movie content rather than TV shows.

According to the countplot, it appears that Netflix adds the highest number of movies and TV shows during the period between October and January. This period seems to be the busiest time of year for Netflix in terms of adding new content to its platform. Netflix has the highest number of content in the United States, followed by India. 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 visualized using a dendrogram.

Finally, a content-based recommender system was created using the similarity matrix obtained through cosine similarity. This system provides personalized 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 personalized recommendations based on the user's viewing history. This study provides valuable insights into the factors influencing the popularity of movies and TV shows on Netflix, offering a foundation for further research and analysis.

Contributors Roles:

Sandeep Salunke

1. Data Loading:
2. Data handling
3. Handling missing values
4. Data exploration/Visualization
5. Outliers detection
6. Hypothesis Testing
7. Feature engineering
8. Model deployment

Github Repo link

https://github.com/Sandeep81299/Netflix-Movies-and-TV-shows-Prediction

to each other. That means more the people more will be adr. is_repeated guest and previous bookings not canceled has strong correlation. may be repeated guests are not more likely to cancel their bookings.