Capstone Project Submission

Name, Email and Contribution:

Name- Potdar Vinayak
Email- vinayak.potdar.vp@gmail.com

Contributions:
Data Wrangling
Null Value Analysis and Treatment
Exploratory Data Analysis
Data Preprocessing and Feature Selection
Model fitting and Optimal Cluster analysis

Please paste the GitHub Repo link.

Github Link:- https://github.com/vinayakpotdar2114/Capstone-Project-4-Netflix-Movies-and-TV-Shows-Clustering

Please write a short summary of your Capstone project and its components. Describe the problem statement, your approaches and your conclusions. (200-400 words)

Problem Statement and Data Description
In this project, I was required to do

1. Exploratory Data Analysis
2. Understanding what type content is available in different countries
3. Is Netflix increasingly focusing on TV rather than movies in recent years.
4. Clustering similar content by matching text-based features.

Approach
The data was in a really good shape, I started off by understanding the data followed by Exploratory Data Analysis. After the EDA, I split the data into two parts, one being the text-based data and the other with the remaining data.

The text-based data was used for 'clustering model' fitting. This required the data to be in a vectorized form and cleaned. This was achieved by removing stop words, punctuations, and stemming. The vectorization technique used was 'Tf-idf'. In order to find the hidden themes of the data, I performed LDA and LSA where I extracted 10 topics.

Now, I needed to get the optimal number of clusters from the dataset for which I used Silhouette Score Analysis , Elbow Method and Dendrogram. The optimal number of clusters turned out to be **3**.

For getting the clusters formed using the vectors of the text-based data, KMeans Model is a tried and tested option. I went ahead with Kmeans and achieved 3 clusters of movies and TV shows. This was followed by again extracting the theme of the clusters by getting the top valued words from each cluster.

Conclusion

We performed EDA and these were the results-

- The platform is increasingly focusing on TV-Shows.
- The platform publishes the most content at the end of the year when the holidays are right around the corner.
- US, India and UK are the top three publishers on the platform.
- There are twice as much Movies as TV shows on the platform, but the increase in the publishing of TV shows can be seen, where the same cannot be stated for movies.
- The top genre of the content in International Movies, Dramas and Comedy.
- Majority of the content on the platform is for mature audience only.

The movies and the tv shows were successfully clustered into three clusters using Kmeans algorithm where the topics of the clusters are—

Cluster 0: drama movi intern comedi independ romant young woman horror love
Cluster 1: tv show intern crime british romant kid seri korean spanishlanguag
Cluster 2: documentari action adventur famili standup children intern movi kid world