Capstone Project 4 Submission

Netflix Movies and TV Shows Clustering

Capstone Project Summary

This Project was of unsupervised type and was aimed at clustering, NLP and EDA of the **Netflix Data** set , to come up with various insights and conclusions & recommendations in order to maximize the customer base and the number of active users watching movies or TV shows .

At first we found out the missing values in the dataset and then after replacing the null values and cleaning the dataset, we performed EDA on it and found out below observations ,

Our Findings :-

- The ratio of TV shows vs Movies is 31: 69 , thus the total number of movies is more than double of number of TV shows & web series ..,
- Most of the Tv shows/web series have max 3 seasons , this refers that the forthcoming Tv shows can be introduced in similar 2-3 series max else it loses its TRP (viewership).
- The Movie's duration is 90 min – 110 mins of most movies , which can be inferred that people are comfortable watching movies within this time duration .
- Maximum number of Movies as well as TV shows are Rated under TV-MA category and then followed by TV-14 .
- There is a trend that most of the movies and TV shows are added during October to January months of the year .
- The number of movies and TV shows added has constantly risen from 2015 to 2019 but we have found out that more number of movies were added during this tenure as compared to TV shows but between 2019- 2020 yes there has been more TV shows added than Movies .
- Then we have tabulate top 15 countries which has realeased most movies and TV shows and found that US is on top in both. In shows US is followed by UK and in movies US is followed by India.
- We also have actors who were part of most shows and movies which gave us that Takahiro sukurai has worked in most shows and Anupam kher has worked in most movies.
- We also found out that Genre wise International movies & Tv shows are most in numbers followed by Drama & Comedy categories and this order is same for Movies and Tv shows.

- Also Using NLP we have found out top 10 recommended movies and Tv shows by implementing Cosine similarity which has most similarity with the other content present on Netflix , using this features the suggested TV shows and Movies can be recommended to a particular user based upon his past preferences . .

Final Conclusion on clustering :-

- Using DB Scan ,As shown by the color codes of the DB Scan graph, it can be seen that the algorithm is clustering the data into 3 main clusters ..,

- Using Agglomerative hierarchical clustering , In our case we got 2 vertical lines, which symbolizes that the given data can be best divided into 2 clusters ..,

- Using K-Means Clustering ,Silhouette score closest to 1 is considered as best and its respective value shown the optimal no. of clusters for the given dataset data.., Here 2 No. of clusters are optimal number of clusters because its respective silhouette score 0.6 was the highest amongst all ..,

- **Spectral Clustering**
- **K –Modes**
- **DB Scan**
- **K- Means**
- **Hierarchical agglomerative**

Conclusion :-

For NetFlix Movies , International Movies with genres as Drama ,Comedy & thriller etc. having duration of 90-110 mins are the most preferred ones , and are the safe bet for good viewership and for attracting more new subscriptions.., also people like to watch actors like Anupam Kher , Shahrukh Khan , Akshay Kumar , Kareena Kapoor , Boman Irani , Paresh Rawal etc.

For NetFlix TV shows , we recommend to introduce international TV shows , followed by Drama , Comedy and Crime related series between October to January period with actors like Takahiro Sakurai , Junichi Suwabe , Yuki Kaji etc , this would be safe & profitable bet as there will be more chance of the show getting hit .,

As far as **Clustering** is concerned of the given dataset ,Some algorithms are forming 3 clusters and majority of them 2 ,so we can conclude as the total clusters formed on our dataset in most of the cases is between 2 clusters .

Team Member’s Name, Email and Contribution:
Contributor Roles :

- **Kundan Lal : kundanlal2001@gmail.com**

- DB Scan
- Hierarchical Agglomerative
- K- Means
- Conclusion
- Summary

- **Pankaj Sudhir Ganjare : ganjarepankaj@gmail.com**

- Outlier detection
- Applying NLP
- Applying algorithms for tv shows

- **Abhijeet Kulkarni : abhijeetkulkarni11020@gmail.com**

- Data preprocessing
- Data cleaning
- Performing EDA
- Data visualization

- **Mohd. Sharik : livisliquoro999@gmail.com**

- K-Modes
- EDA
- PCA
- Spectral Clustering
- Conclusion

Github Link:-
https://github.com/MohdSharik99/NETFLIX_MOVIES_AND_TV_SHOWS_CLUSTERING

Google Drive link:

https://drive.google.com/drive/folders/1rq0o0X0jDLEixyfe0kV7dzsQfH-ltFSn

Google Drive Link :