Capstone Project- 04

NETFLIX MOVIES AND TV SHOWS CLUSTERING

Team Member's Name, Email and Contribution:

Contributor's Role:

Abhishek Kirar

- Introduction
- Data cleaning
- Null Value Treatment
- Exploratory Data Analysis
- Univariate
- Data cleaning & pre-processing for clustering
- Encoding the categorical data
- K means clustering
- Hierarchical Clustering
- Silhouette analysis
- Conclusion

GitHub Repo link:

https://github.com/abkirar27/Netflix_Movies_And_Tv_Shows_Clustering_Capstone_Project-04

Here is the short summary of the Capstone project and its components.

PROBLEM

This dataset consists of Tv-shows and movies available on Netflix as of 2019. The dataset is collected from flixable which is a third-party Netflix search engine.
In 2018, they released an interesting report which shows that the number of TV shows on Netflix has nearly tripled since 2010. The streaming service's number of movies has decreased by more than 2,000 titles since 2010, while its number of TV shows has nearly tripled. It will be interesting to explore what all other insights can be obtained from the same dataset. Integrating, this dataset with other external datasets such as IMDB ratings, rotten tomatoes can also provide many interesting findings.

APPROACH

Initially, in the 1st step imported the data set to carry out the analysis over the data set to comprehend the details of available data and Checked for Null values and treated them. Here, we found more than 30% null values in the director's column. Then, we take appropriate action for null values according to the circumstances.

Performed the Exploratory data analysis and tried to get the understanding of the data and how the content is distributed in the dataset, its type and details such as which countries are watching more and which type of content is in demand etc. has been analyzed in this step with the help of visualization graph by getting insights from analysis.

- Data preprocessing – in this we remove the punctuation and stops words also used stemming to reduce words to their basic form or stem, which may or may not be a legitimate word in the language.
- We used the k-means clustering algorithm and then checked the model performance using Silhouette's coefficient and elbow method to find the number of clusters.

Analyzing all the variables of the data set and identifying the solution for given tasks.

Performed hypothesis testing to get the insights on duration of movies and content with respect to different variables.

After doing feature engineering and finding the number of clusters, we used the k-means algorithm and then checked the model performance using Silhouette's coefficient, to identify the best fit Model.

The number of movies on Netflix is growing significantly faster than the number of TV shows. Because of covid-19, there is a significant drop in the number of movies and television episodes produced after 2019.

- The project's main goal is to create a model that can perform Clustering on comparable material by matching text-based attributes.
- As the problem statement says, understanding what type of content is available in different countries and Is Netflix increasingly focused on TV rather than movies in recent years we have to do clustering on similar content by matching text-based features. For that we used Affinity Propagation, Agglomerative Clustering, and K-means Clustering.
- In Affinity Propagation, we had 13 clusters and a Silhouette Coefficient score

of 0.244.

- In Agglomerative Clustering. There were four clusters, with an average silhouette score of 0.17296314851287742.

The final model we used wask-means clustering, which consisted of 2,3,4,5,6 clusters.

- For n_clusters = 2 The average silhouette_score is : 0.7049787496083262
- For n_clusters = 3 The average silhouette_score is : 0.5882004012129721
- For n_clusters = 4 The average silhouette_score is : 0.6505186632729437
- For n_clusters = 5 The average silhouette_score is : 0.56376469026194
- For n_clusters = 6 The average silhouette_score is : 0.4504666294372765

CONCLUSION

- Data set contains 7787 rows and 12 columns in that cast and director features contains large number of missing values so we can drop it and we have 10 features for the further implementation.
- We have two types of content TV shows and Movies (30.86% contains TV shows and 69.14%contains Movies).
- The most number of the movies and TV shows release in 2017 and 2020 respectively and united nation have the maximum content on Netflix
- On Netflix, Dramas genre contains the maximum content among all of the genres and the most of the content added in December month and less content in February.
- By applying the silhouette score method for n range clusters on dataset we got best score which is 0.244for 4 clusters it means content explained well on their own clusters, by using elbow method after k =4 curve gets linear it means k =4 will be the best cluster
- By applying different clustering algorithms to our dataset, we get the optimal number of cluster is equal to 4.
- We have done null value treatment, feature engineering, and EDA since loading the dataset then completed assigned tasks.
- Among different types of content available in different countries, content TV-MA is available in most countries. This could be because it shows that it is just for adult audiences, and the Netflix audience enjoys content like this.
- We have also explained different clusters based on their content; Defined clusters and enforced the K-means clustering algorithm and cluster number nine has the most clusters; we have also plotted a scatter plot in which we may interact with similar content about that cluster.