CAPSTONE PROJECT 4

Clustering Model

Netflix Movies and Tvshows Clutering

Prepared by:

PRAVEEN .S
(Cohort Everest)

Data Science Trainee

Almabetter

Abstract

Netflix, Inc. is an American subscription streaming service and production company. Launched on August 29, 1997, it offers a film and television series library through distribution deals as well as its own productions, known as Netflix Originals.As of March 31, 2022, Netflix had over 221.6 million subscribers worldwide.

Our clustering model can help us understand what number of and type of contents available in this streaming platform and its success behind by identifying the number of clusters of contents on the basis of its parameters.

Problem Statement:

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.

This dataset consists of about **7500+** records and **12** attributes

1.**show_id :** Unique ID for every Movie / Tv Show
2**.type :** Identifier - A Movie or TV Show
3.**title :** Title of the Movie / Tv Show
4.**director :** Director of the Movie
5.**cast :** Actors involved in the movie / show
6.**country :** Country where the movie / show was produced
7.**date_added :** Date it was added on Netflix
8.**release_year :** Actual Releaseyear of the movie / show
9.**rating :** TV Rating of the movie / show
10**.duration :** Total Duration - in minutes or number of seasons
11.**listed_in** : Genere
12**.description:** The Summary description

Introduction:

I had presented my exploratory analysis,,visualization,Correlationplots and developed three machine learning clustering models and lot of other interesting insights into the given dataset.I choose this particular dataset as this is the domain we had day to day experience as the world is revolving around streaming platforms.My goal here is to build a clustering model, which couldhelp us to interfere with the type of contents available in streaming platforms as such.

Steps involved:

Exploratory Data Analysis

After loading the dataset we EDA by comparing all the variables to get a proper outlook of the dataset and also dropped the columns that is irrelevant in order to obtain a cluster model, thus It gave us a better idea of which feature behaves in which manner and done the following steps to do data visualization.

Null values Treatment

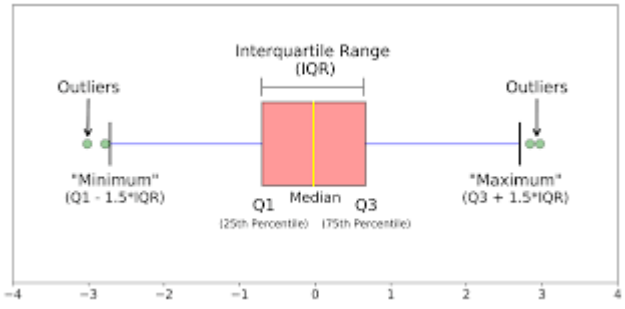
After finding that the dataset have about 3000+ missing values which was found using the missingno plot, we have replaced most of the values with similar context and have dropped the rows which has more missing values to obtain a ideal dataset. We interpret the rating using a sub barplot of each movies and shows.We found some countries are producing TV shows than movies.Also we found that the contents uploaded are increasing significantly using a lineplot

Data Visualization:

In the visualization, we found that the distribution of movies and Tv shows using a donut plot and which country most of the contents using barplot. Using a distplot we came to observe the distribution of movie duration.We interpret the rating using a sub barplot of each movies and shows.We found some countries are producing TV shows than movies.Also we found that the contents uploaded are increasing significantly using a lineplot.The description of the contents has a average of 140 word which we found using a boxplot.Finally the most used words in the description are observed using a wordcloud.

Outlier Detection:

Dropped the TV shows that has more than 9 seasons as they are very less in number.And also using the quantile range method dropped the movies that has more than 250 mins and less than 50 mins



Data Preparation:

From NLTK we have imported stop word to remove common words to do natural Language Processing and have found the most used words in the description by plotting a word cloud .Defined a function to remove punctuation in the description column and also stemed each word using the lemmatizer function.And Finally converted the clustered data-frame into an array for modeling.

Tokenization:

Tokenization is essentially splitting a phrase, sentence, paragraph, or an entire text document into smaller units, such as individual words or terms. Each of these smaller units are called tokens. The tokens could be words, numbers or punctuation marks**.**

TfidfVectorizer

In TfidfVectorizer we consider overall document weightage of a word. It helps us in dealing with most frequent words. Using it we can penalize them. TfidfVectorizer weights the word counts by a measure of how often they appear in the documents

Count Vectorizer

	blue	bright	sky	sun
Doc1	1	0	1	0
Doc2	0	1	0	1

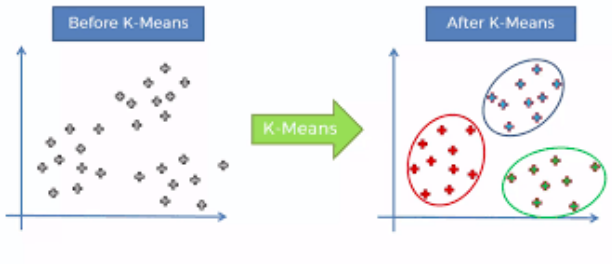
TD-IDF Vectorizer

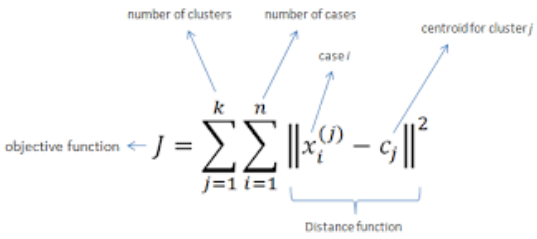
	blue	bright	sky	sun
Doc1	0.707107	0.000000	0.707107	0.000000
Doc2	0.000000	0.707107	0.000000	0.707107

Algorithms for Modeling:

K- Means Clustering:

The K-means clustering algorithm computes centroids and repeats until the optimal centroid is found. It is presumptively known how many clusters there are. It is also known as the flat clustering algorithm. The number of clusters found from data by the method is denoted by the letter ‘K’ in K-means. In this method, data points are assigned to clusters in such a way that the sum of the squared distances between the data points and the centroid is as small as possible. It is essential to note that reduced diversity within clusters leads to more identical data points within the same cluster.

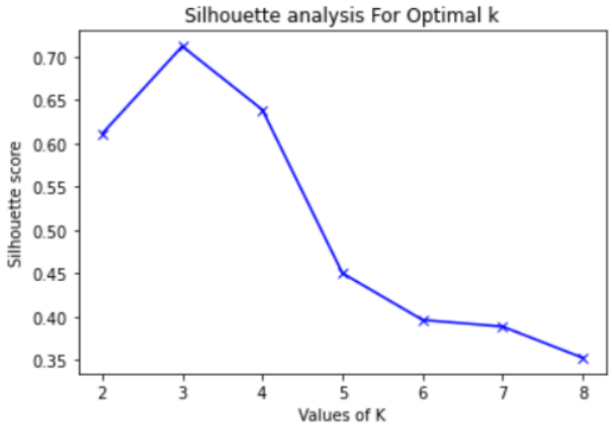




Silhouette score:

Silhouette score is used to evaluate the quality of clusters created using clustering algorithms such as K-Means in terms of how well samples are clustered with other samples that are similar to each other. The Silhouette score is calculated for each sample of different clusters

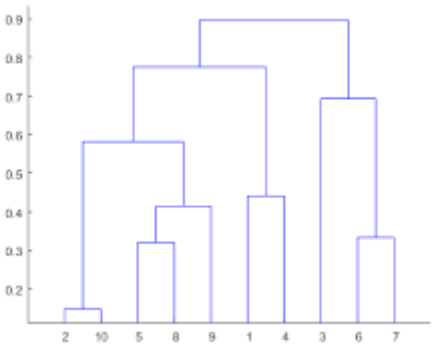
$$Silhouette - score = \frac{b_i - a_i}{\max(bi, a_i)}$$



Line plot between K and Silhouette score

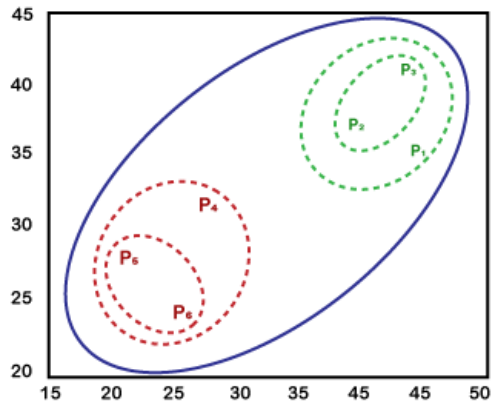
Dendrogram:

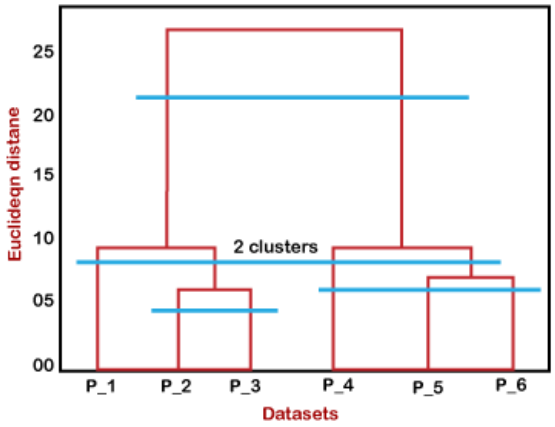
A dendrogram is a diagram that shows the hierarchical relationship between objects. It is most commonly created as an output from hierarchical clustering. The main use of a dendrogram is to work out the best way to allocate objects to clusters.



Hierarchial Clustering

Hierarchical clustering involves creating clusters that have a predetermined ordering from top to bottom. For example, all files and folders on the hard disk are organized in a hierarchy. There are two types of hierarchical clustering, Divisive and Agglomerative.





8. Conclusion:

That's it! We reached the end of our exercise.

Starting with loading the data so far we have done EDA , null values treatment,Outlier Detection,Data Preprocessing by removing punctuation marks in the Description, and then model building. In all of these clustering models estimated the optimal no.of clusters to be 6 which is also validated by the sillhouette score. We also done a hierarchical Clustering to this dataset as this became relevant when using the Dendrogram.
Finally, we found that the optimal no. Of clusters to be 6.

References:

1. MachineLearningMastery
2. GeeksforGeeks
3. Analytics Vidhya
4. Wikipedia