**Netflix Movies and TV Shows Clustering**

**Shubham Naik**

**Data science trainee,**

**Almabetter, Bangalore**

**Abstract**

Netflix is a subscription-based streaming service that allows their members to watch TV shows and movies without commercials on an internet-connected device. User can also download TV shows and movies to their iOS, Android, or Windows 10 device and watch without an internet connection. However, user can cancel their subscriptions at any time. Therefore, the company must keep the users engage on the platform and not lose their interest. This is where recommendation systems start to play an important role, providing valuable suggestions to users is essential.

Recommendation system aimed to predict the number of stars that a user would rate a title, on a scale of one to five, after they had watched it. Once a user returned a title, they provided their actual rating. This served as the primary source of feedback from the user that the company then used to retrain and optimize its recommendation algorithms.

***Keywords:*** *Machine Learning****,*** *Clustering, Recommendation*

**Introduction**

Netflix is an American digital content streaming and production company that was founded in 1997 by Reed Hastings and Marc Randolph. The company initially offered a mail-based DVD rental subscription service, but it introduced digital video streaming services in 2007. In 2016, the company separated its DVD service onto a new platform, known as DVD.com: A Netflix Company. The company then transitioned its main platform to subscription-based digital video streaming. The company is regarded as the world’s largest subscription streaming service, with most recent estimates suggesting that the company now has 167.1 million subscribers around the world, with over 100 million of these users residing outside the United States. Netflix is streaming in more than 30 languages and 190 countries and territories.

Recommendation system is an important contributor to its revenue generation model, driving approximately 80 percent of hours of content streamed on the platform. Unlike YouTube and Amazon, the platform does not deliver targeted advertisements to its users. Rather, the company relies on subscriptions to both its digital video streaming service and DVD-delivery service to generate revenue.  Recommendation system holds a significant amount of influence over how the platform operates and how users engage with the service.

**Problem Statement**

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.

**Attribute Information**

* show\_id: Unique ID for every Movie / Tv Show
* type: Identifier - A Movie or TV Show
* title: Title of the Movie / Tv Show
* director: Director of the Movie
* cast: Actors involved in the movie / show
* country: Country where the movie / show was produced
* date\_added: Date it was added on Netflix
* release\_year: Actual Release year of the movie / show
* rating: TV Rating of the movie / show
* duration: Total Duration - in minutes or number of seasons
* listed\_in: Genre
* description: The Summary description

**Steps Involved**

1. **Handling Missing Values**

Missing data is defined as the values or data that is not stored (or not present) for some variable/s in the given dataset. In our dataset there were 5 columns which has missing values director, cast, country, date\_added and rating. Since date\_added and rating have very few null values we can drop their row of total 17 observations. It wont affect data significantly. In director, cast and country we substitute the null values with the word ‘No Data’ for further analysis.

1. **Exploratory Data Analysis**

The main goal of EDA is to understand the data and represent in such a way that everything given in the data makes sense. So far, we’ve got to know about which data has been given to us and also, we’ve found no missing values in it. So now what? The answer is quite simple, we now analyze what are the key information we can gather from the given data, which not only helps us to better understand about the data but also it is quite simple for any non-technical person to understand it too. For now, we will look at the EDA part.

Explorations and visualizations are as follows

1. Distribution of type
2. Rating and shows of movies
3. Relation between type and rating
4. Country wise content production
5. Top 5 countries separated by type of release
6. Content release over the years
7. Content addend over the years
8. Day wise addition of movies and TV shows to the platform
9. Month wise addition of movies and TV shows to the platform
10. Date wise addition of movies and TV shows to the platform
11. Movies that take less amount of time
12. Duration of Tv shows
13. Duration of movies
14. Top 25 Directors
15. Genres for movies
16. Genres for Tv shows
17. Most used words in titles
18. **Hypothesis Testing**

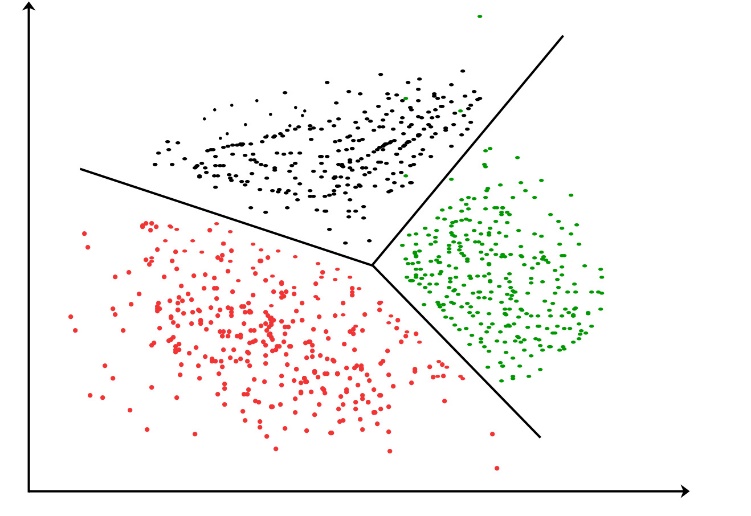
Hypothesis testing is a common statistical tool used in research and data science to support the certainty of findings. The aim of testing is to answer how probable an apparent effect is detected by chance given a random data sample. A hypothesis is often described as an “educated guess” about a specific parameter or population. Once it is defined, one can collect data to determine whether it provides enough evidence that the hypothesis is true.

In hypothesis testing, two mutually exclusive statements about a parameter or population (hypotheses) are evaluated to decide which statement is best supported by sample data.

1. **Clustering**

Clustering is a Machine Learning technique that involves the grouping of data points. Given a set of data points, we can use a clustering algorithm to classify each data point into a specific group. In theory, data points that are in the same group should have similar properties and/or features, while data points in different groups should have highly dissimilar properties and/or features. Clustering is a method of unsupervised learning and is a common technique for statistical data analysis used in many fields.

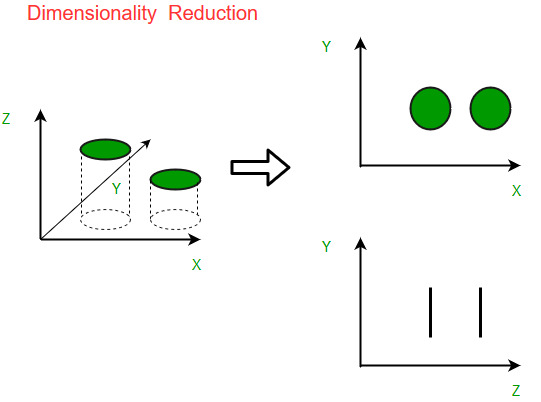
K-Means is probably the most well-known clustering algorithm. Each data point is classified by computing the distance between that point and each group center, and then classifying the point to be in the group whose center is closest to it.



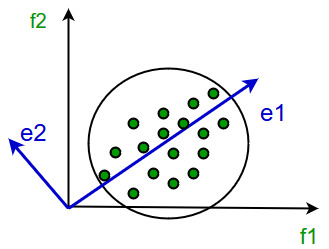
1. **Dimensionality Reduction**

Dimensionality reduction is the process of reducing the number of random variables under consideration, by obtaining a set of principal variables. It can be divided into feature selection and feature extraction.

A classification problem that relies on both humidity and rainfall can be collapsed into just one underlying feature, since both of the aforementioned are correlated to a high degree. Hence, we can reduce the number of features in such problems. A 3-D classification problem can be hard to visualize, whereas a 2-D one can be mapped to a simple 2-dimensional space, and a 1-D problem to a simple line. The below figure illustrates this concept, where a 3-D feature space is split into two 1-D feature spaces, and later, if found to be correlated, the number of features can be reduced even further.



There are various methods for dimensionality reduction in our project we used Principal Component Analysis (PCA). This method was introduced by Karl Pearson. It works on a condition that while the data in a higher dimensional space is mapped to data in a lower dimension space, the variance of the data in the lower dimensional space should be maximum.



**It involves the following steps:**

* Construct the covariance matrix of the data.
* Compute the eigenvectors of this matrix.
* Eigenvectors corresponding to the largest eigenvalues are used to reconstruct a large fraction of variance of the original data.

Hence, we are left with a lesser number of eigenvectors, and there might have been some data loss in the process. But, the most important variances should be retained by the remaining eigenvectors.

1. **Data Preprocessing:**

**Removing Stopwords:** The process of converting data to something a computer can understand is referred to as pre-processing. One of the major forms of pre-processing is to filter out useless data. In natural language processing, useless words (data), are referred to as stop words. A stop word is a commonly used word (such as “the”, “a”, “an”, “in”) that a search engine has been programmed to ignore, both when indexing entries for searching and when retrieving them as the result of a search query.

We would not want these words to take up space in our database, or taking up valuable processing time. For this, we can remove them easily, by storing a list of words that you consider to stop words.

**Removing Punctuation:** Punctuations does not carry any meaning in clustering, so removing punctuations helps to get rid of unhelpful parts of the data, or noise.

1. **K-means Clustering**

K-means clustering is one of the simplest and popular unsupervised machine learning algorithms. The objective of K-means is simple: group similar data points together and discover underlying patterns. To achieve this objective, K-means looks for a fixed number (k) of clusters in a dataset. A cluster refers to a collection of data points aggregated together because of certain similarities. You’ll define a target number k, which refers to the number of centroids you need in the dataset. A centroid is the imaginary or real location representing the center of the cluster. Every data point is allocated to each of the clusters through reducing the in-cluster sum of squares. In other words, the K-means algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. The ‘means’ in the K-means refers to averaging of the data; that is, finding the centroid.

**How the K-means algorithm works:**

To process the learning data, the K-means algorithm in data mining starts with a first group of randomly selected centroids, which are used as the beginning points for every cluster, and then performs iterative (repetitive) calculations to optimize the positions of the centroids

It halts creating and optimizing clusters when either:

* The centroids have stabilized — there is no change in their values because the clustering has been successful.
* The defined number of iterations has been achieved.

**Silhouette Analysis in K-means Clustering:**

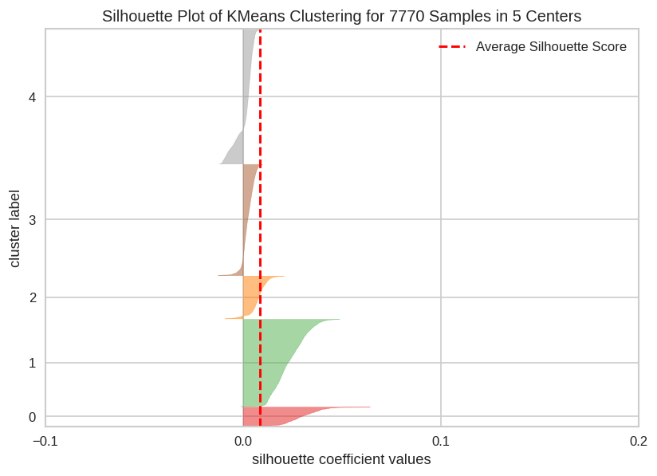
Silhouette score is used to evaluate the quality of clusters created using clustering algorithms such as K-Means in terms of how well samples are clustered with other samples that are similar to each other. The Silhouette score is calculated for each sample of different clusters. To calculate the Silhouette score for each observation/data point, the following distances need to be found out for each observation belonging to all the clusters:

* Mean distance between the observation and all other data points in the same cluster. This distance can also be called a **mean intra-cluster distance.**The mean distance is denoted by **a**
* Mean distance between the observation and all other data points of the next nearest cluster. This distance can also be called a **mean nearest-cluster distance.** The mean distance is denoted by **b**

Silhouette score, **S,** for each sample is calculated using the following formula:

\(S = \frac{(b - a)}{max(a, b)}\)

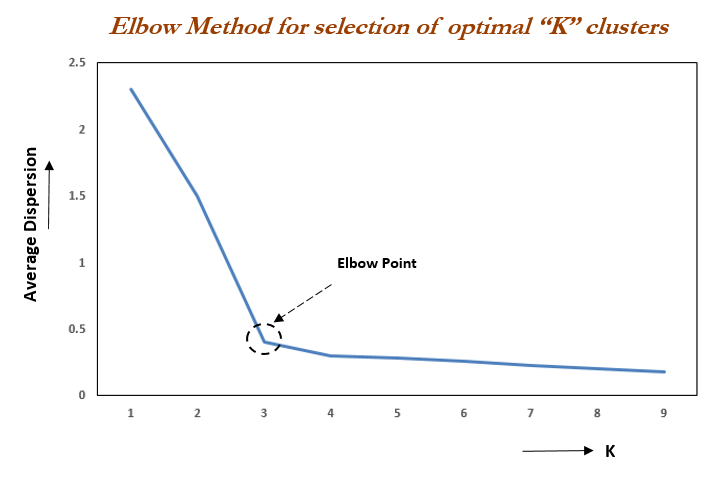
The value of the Silhouette score varies from -1 to 1. If the score is 1, the cluster is dense and well-separated than other clusters. A value near 0 represents overlapping clusters with samples very close to the decision boundary of the neighboring clusters. A negative score [-1, 0] indicates that the samples might have got assigned to the wrong clusters.



**Elbow curve method**

The elbow method is used to determine the optimal number of clusters in k-means clustering. The elbow method plots the value of the cost function produced by different values of k. As you know, if k increases, average distortion will decrease, each cluster will have fewer constituent instances, and the instances will be closer to their respective centroids. However, the improvements in average distortion will decline as k increases. The value of k at which improvement in distortion declines the most is called the elbow, at which we should stop dividing the data into further clusters.

When we plot the graph of ‘value of k’ on x-axis and ‘value of Epsilon’ on y-axis, there is an elbow formation at the optimum value of ‘k’.

****

From the above graph we observe that there is an elbow formation at k = 3. Hence the optimum value of k is 3. Therefor we cluster the data set into 3 clusters.

1. **TF-IDF Vectorizer**

TF-IDF is an abbreviation for Term Frequency Inverse Document Frequency. This is very common algorithm to transform text into a meaningful representation of numbers which is used to fit machine algorithm for prediction.

**Term Frequency (tf)**

tf is the number of times a term appears in a particular document. So, it’s specific to a document. A few of the ways to calculate tf is given below: -

tf(t) = No. of times term ‘t’ occurs in a document

**OR**

tf(t) = (No. of times term ‘t’ occurs in a document) / (No. Of terms in a document)

**Inverse Document Frequency (idf)**

idf is a measure of how common or rare a term is across the entire corpus of documents. So the point to note is that it’s common to all the documents. If the word is common and appears in many documents, the idf value (normalized) will approach 0 or else approach 1 if it’s rare. A few of the ways we can calculate idf value for a term is given below

idf (t) =1 + log e [ n / df(t)]

**OR**

idf(t) = log e [ n / df(t)] where

n = Total number of documents available

t = term for which idf value has to be calculated

df(t) = Number of documents in which the term t appears

1. **Cosine Similarity**

similarity measure refers to distance with dimensions representing features of the data object, in a dataset. If this distance is less, there will be a high degree of similarity, but when the distance is large, there will be a low degree of similarity.

**Some of the popular similarity measures are** –

* Euclidean Distance.
* Manhattan Distance.
* Jaccard Similarity.
* Minkowski Distance.
* Cosine Similarity.

Cosine similarity is a metric, helpful in determining, how similar the data objects are irrespective of their size. We can measure the similarity between two sentences in Python using Cosine Similarity. In cosine similarity, data objects in a dataset are treated as a vector. The formula to find the cosine similarity between two vectors is –

Cos (x, y) = x . y / ||x|| \* ||y||

where,

x . y = product (dot) of the vectors ‘x’ and ‘y’.

||x|| and ||y|| = length of the two vectors ‘x’ and ‘y’.

||x|| \* ||y|| = cross product of the two vectors ‘x’ and ‘y’.

1. **Conclusion**

* Nearly 70 percent are movies and 30 percent are TV Shows
* After 2016 Movies and Tv Shows added maximum
* There are two different type of time durations for Movies it's in minutes and for TV Shows it's in season
* Maximum movies are in the range of 90 to 120 minutes
* Most of the children, sci-fi & Fantasy movies and documentaries take less amount of time
* K-Means Clustering with silhouette gives the highest score of 82% for number of clusters 5

**References:**

1. O’REILLY
2. Towards Data Science
3. Wikipedia
4. Analytics Vidhya
5. Medium
6. GeeksforGeeks