STEVENS
INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY®

Recommender System For Netflix

BIA 686- A

Final Project

Submitted By,

Ruchi Hiralal Mendhegiri

Index

- Netflix Background
- Recommender System
- Business Impact
- Data Overview and Attributes
- Motivation and Problem Statement
- Methodology
  1. Data Collection and Overview
  2. Data Cleaning and Observations
  3. Exploratory Data Analysis
  4. Recommender System Development
     - Content Based
     - Collaborative
  5. Evaluation
- Conclusion
- Data Source

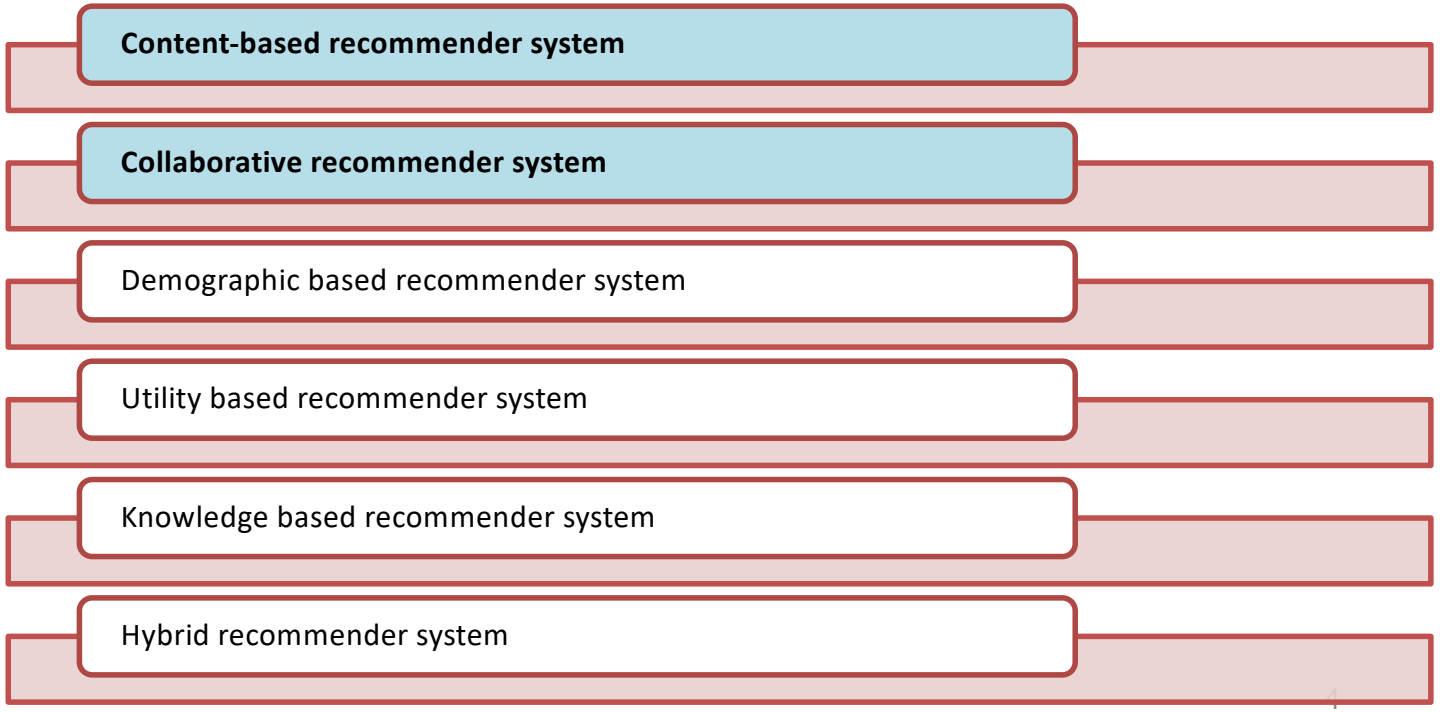
Netflix Background

- Netflix is an American technology and media service provider and production company
- Founders: Marc Randolph and Reed Hastings (1997)
- Corporate headquarters are in Los Gatos, California
- Offers subscription-based Video on demand from a library of films and television series
- Netflix is the largest entertainment/media company by market capitalization
- As of October 19, 2021, Netflix had 214 million subscribers

Ref: https://en.wikipedia.org/wiki/Netflix
https://help.netflix.com/en/node/100639

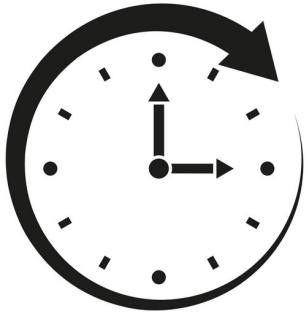
Recommender System

- **Goal: Provide personalized recommendations to guide the user in exploring a large space of possible options.**
- Types of recommender systems :



Ref: https://help.netflix.com/en/node/100639
https://towardsdatascience.com/deep-dive-into-netflixs-recommender-system- 341806ae3b48

Business Impact



Effective recommender system can lead to increased stream time (Close to 80%)

Recommender algorithm boosts retention rate and enhance user experience which translates into savings on customer acquisition costs($1 billion/year)

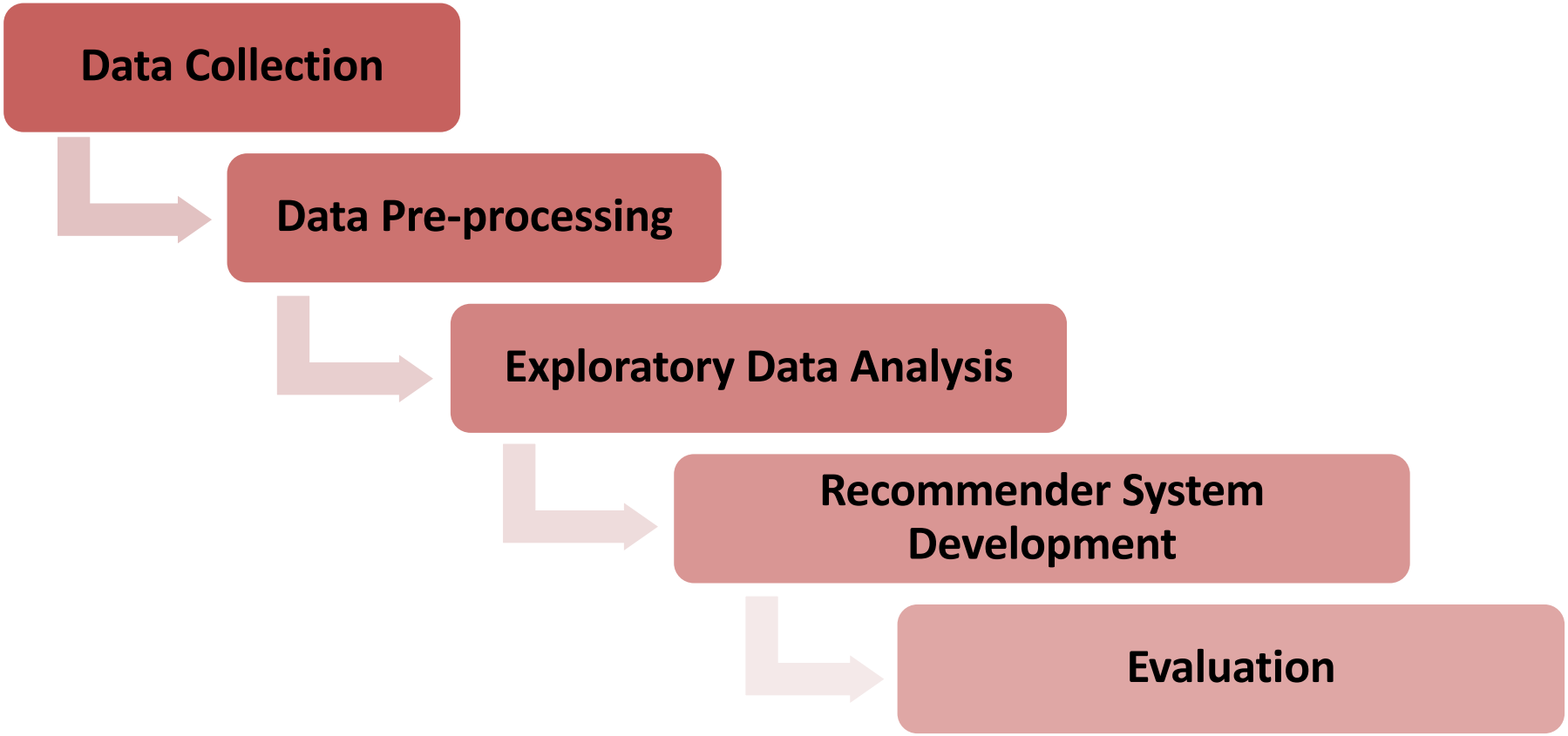
Ref: how-advanced-data-analytics-helped-netflix-generate-billions
https://seleritysas.com/blog/2019/04/05/how-netflix-used-big-data-and-analytics-to- generate-billions/
https://en.wikipedia.org/wiki/Netflix

Motivation and Problem Statement

Business Objective	Filtering Techniques:	Focus Area of this Analysis
• Maximize user satisfaction and user subscription retention, which correlates well with maximizing consumption of video content. • Provide top five recommendations based on customer preferences, streaming patterns and available content	• Content filtering : similarity-based on features (cast, director, country, rating, and genres) • Collaborative filtering : similarity based on ratings given by users	• Understand the dynamics of Netflix dataset • We will deep dive into Netflix data and models that will help to meet our objective and discuss our approach to find best fit model in this space

Ref: https://www.bluepiit.com/blog/classifying-recommender-systems/
https://medium.com/@springboard_ind/how-netflixs-recommendation-engine-works- bd1ee381bf81

Methodology



This project will explore the Netflix dataset through visualizations and graphs using python libraries, matplotlib, and seaborn.

Data Overview and Attributes

- The first dataset consists of TV Shows and Movies available on Netflix as of 2019
- The dataset contains over 6234 titles with 12 attributes
- Attributes Column Names: Show_id, type, title, director, cast, country, date added, release year, rating, duration, listed in, description
- We can also see that there are NaN values in some columns

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	NaN	United States	September 25, 2021	2020	PG-13	90 min	Documentaries	As her father nears the end of his life, filmm...
1	s2	TV Show	Blood & Water	NaN	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	September 24, 2021	2021	TV-MA	2 Seasons	International TV Shows, TV Dramas, TV Mysteries	After crossing paths at a party, a Cape Town t...
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	NaN	September 24, 2021	2021	TV-MA	1 Season	Crime TV Shows, International TV Shows, TV Act...	To protect his family from a powerful drug lor...
3	s4	TV Show	Jailbirds New Orleans	NaN	NaN	NaN	September 24, 2021	2021	TV-MA	1 Season	Docuseries, Reality TV	Feuds, flirtations and toilet talk go down amo...
4	s5	TV Show	Kota Factory	NaN	Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...	India	September 24, 2021	2021	TV-MA	2 Seasons	International TV Shows, Romantic TV Shows, TV ...	In a city of coaching centers known to train I...

Dataset Source : https://www.kaggle.com/shivamb/netflix-shows

Dataset 1 Attributes Definitions

Data Attribute	Definition
SHOW-ID	Unique id of each show
TYPE	Show category. Could be either a Movie or a TV Show
TITLE	Name of the show
DIRECTOR	Name of the director(s) of the show
CAST	Names of Actors/ Actresses in the show
COUNTRY	Countries where the show is available to watch on Netflix
DATE ADDED	Date when the show was added on Netflix
RATING	Show rating on Netflix
RELEASE YEAR	Release year of the show
DURATION	Time duration of the show
LISTED IN	Genre of the show
DESCRIPTION	Brief insight into what the show is about

Dataset 2 and Attribute Definitions

- The second dataset which will be used for collaborative filtering recommender system come directly from Netflix.
- It consists of 4 text data files, each file contains over 20M rows, i.e. over 4K movies and 400K customers. All together over 17K movies and 500K+ customers.
- There are no null values in any columns.

	movie	user	rating	date
0	1	1488844	3	2005-09-06
1	1	822109	5	2005-05-13
2	1	885013	4	2005-10-19
3	1	30878	4	2005-12-26
4	1	823519	3	2004-05-03

Data Attribute	Definition
MOVIE-ID	Unique ID for each new movie record / file
CUSTOMER ID	Unique ID for each new user
RATING	Rating given by user in the range 1 to 5
DATE	Date user gave the ratings

Dataset Source : https://www.kaggle.com/netflix-inc/netflix-prize-data

Data Cleaning and Observations

- There are a total of 4,307 null values across the entire first dataset
- To avoid loss of information, we will use **imputation** treatment
- Director, cast and country has significant amount of missing values:
  - We chose to treat each of these column with missing value as unavailable
- Date added, rating and duration has insignificant missing values
  - We will drop these rows with missing values
- There are no null values in case of second dataset

```
show_id          0
type             0
title            0
director      2634
cast           825
country        831
date_added      10
release_year     0
rating           4
duration         3
listed_in        0
description      0
dtype: int64
```

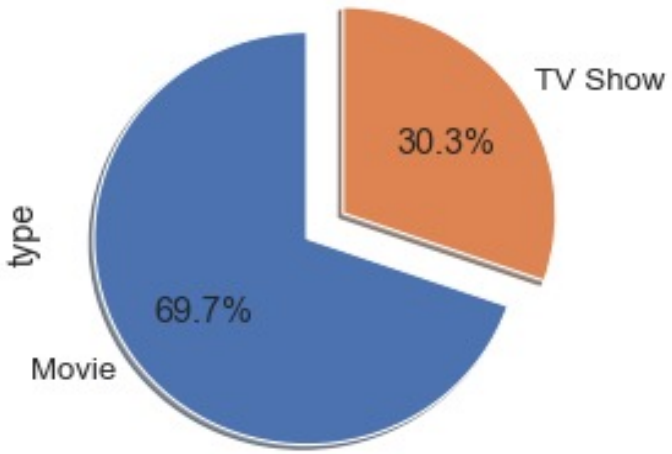
```
movie     0
user      0
rating    0
date      0
dtype: int64
```

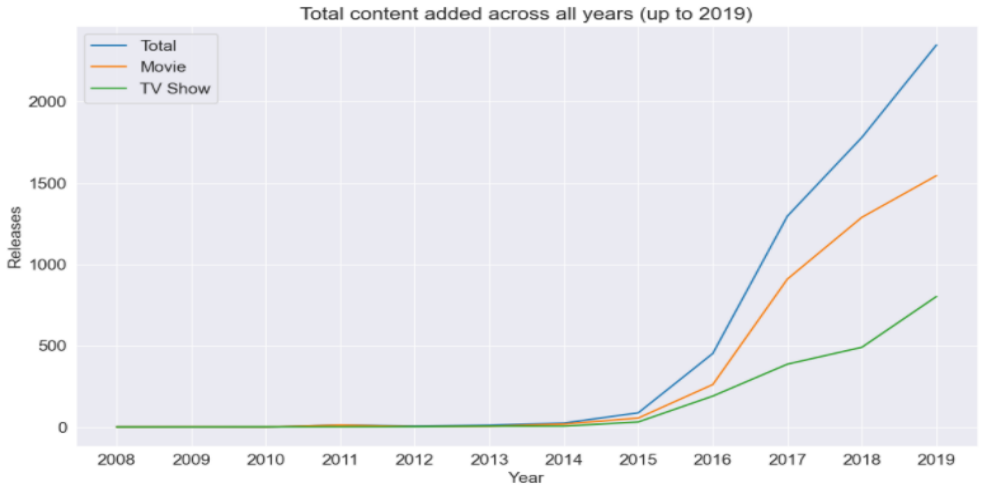
Exploratory Data Analysis

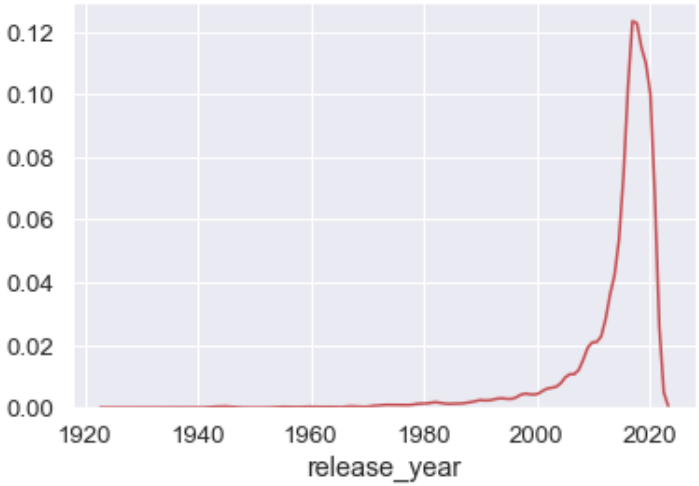
Focus Area of this Analysis :

- Does Netflix have more focus on TV Shows than movies in recent years
- Understanding what content is available in different countries
- What are the popular Genre Content on Netflix
- Actors and Directors ruling Netflix based on amount of Content
- Ratings and Content distribution on Netflix
- Understand user given Rating trends for the movies

1. Does Netflix has more focus on TV Shows than movies in recent years

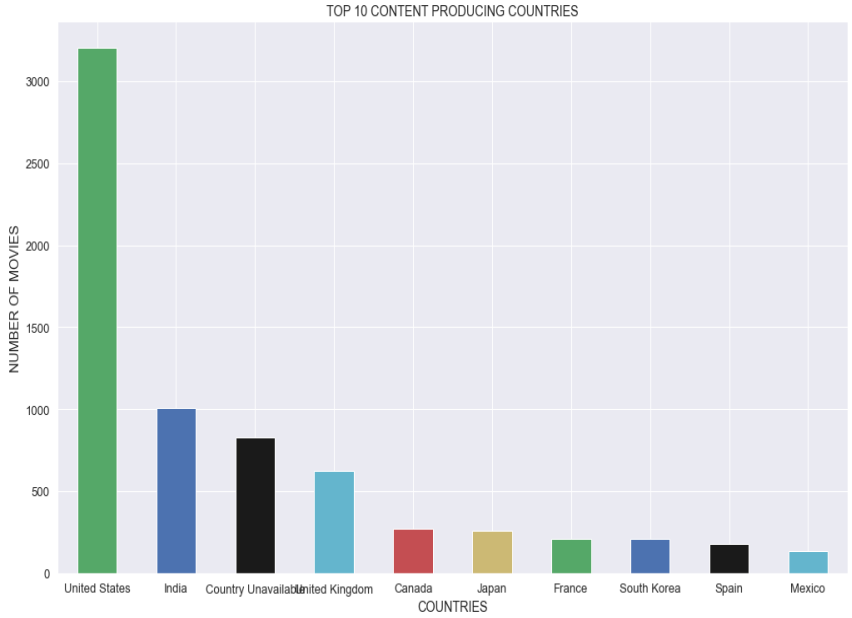


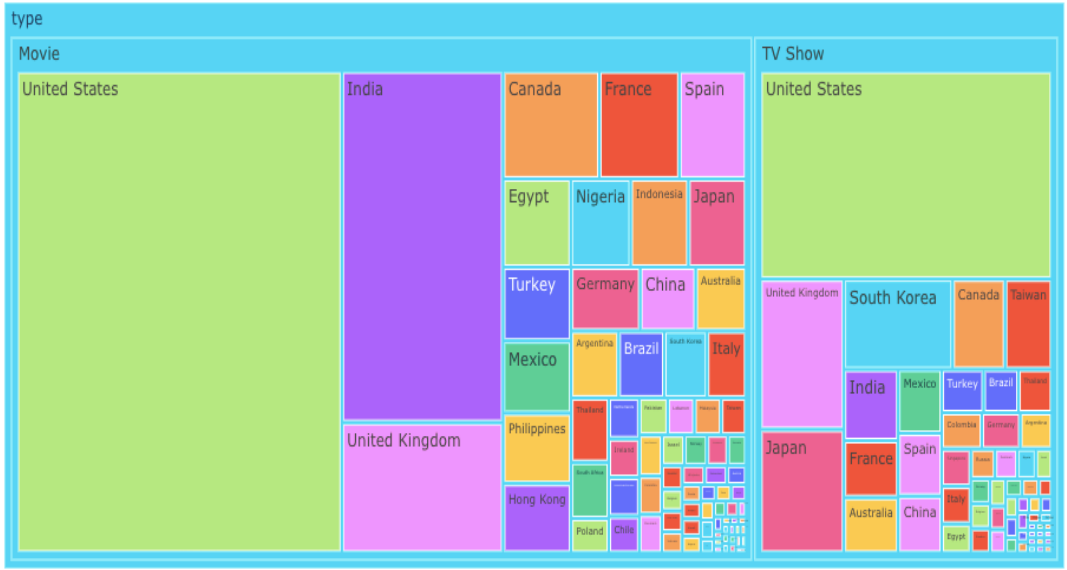




- Netflix dataset consisting of both movies and shows
- It has approximately 69.7% Movie and 30.3% TV shows content
- Maximum content released post 2000
- The popular streaming platform started gaining traction after 2014
- The growth in the number of Movies on Netflix is much higher than that on TV shows

2. Understanding what content is available in different countries

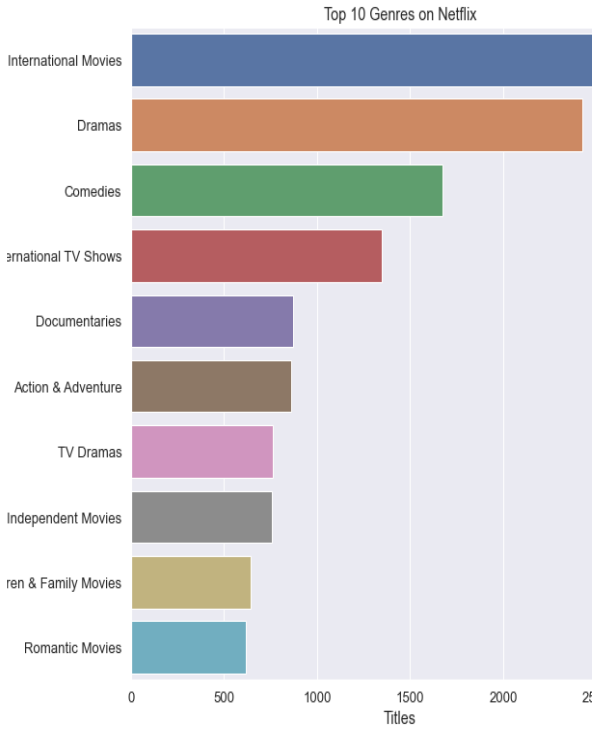


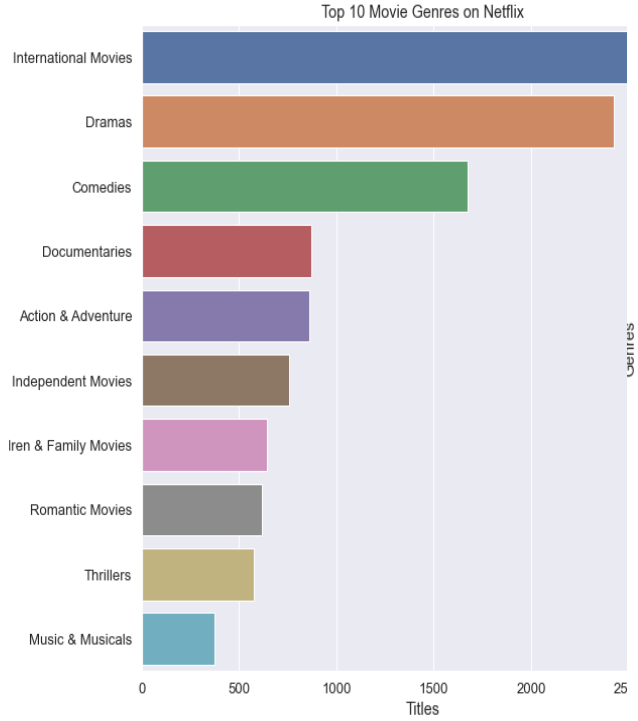


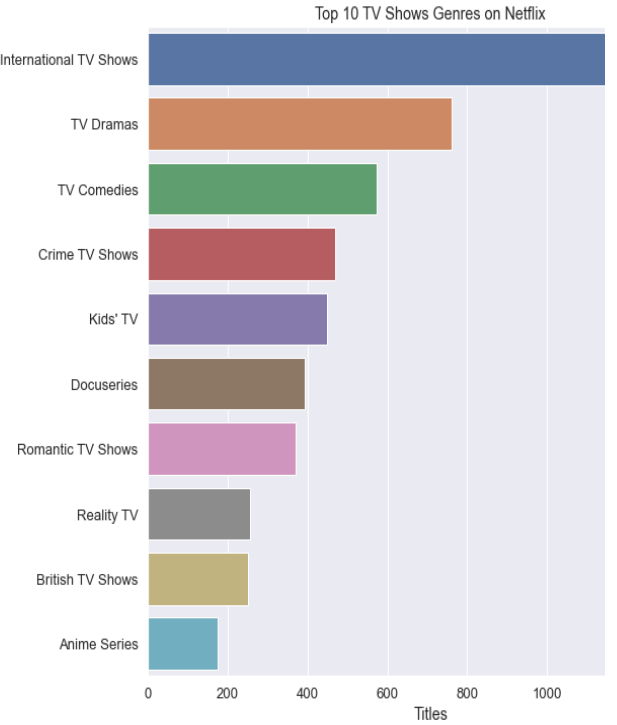
- United States is the first with contents doubling that of India who is second
- United States is at the top for both Movies and Tv Shows
- India is second in terms of number of movies followed by UK & Canada
- United Kingdom is second for tv shows followed by Japan & South Korea

3. Popular Genre Content On Netflix

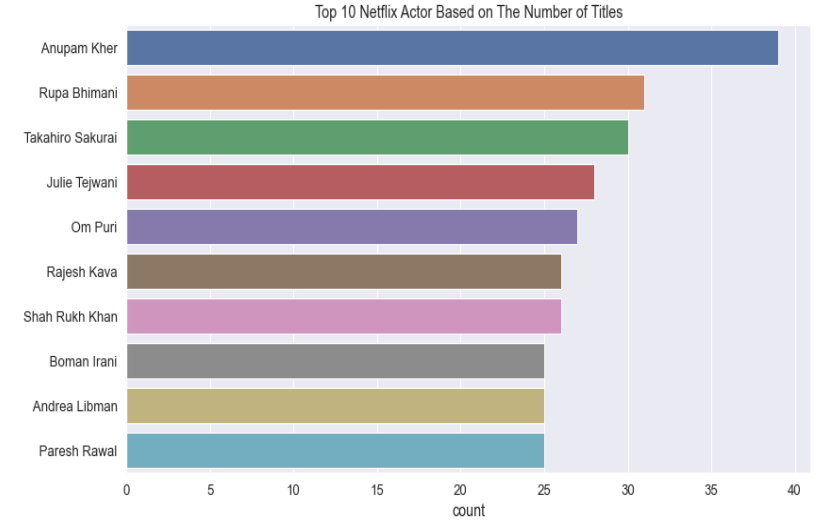
- Overall International Movies take the first place(2500+ titles), followed by dramas (2000+ titles) and comedies (1600+ titles)
- Same trend is followed in Movies and TV shows for top 3 genre







4. Actors and Directors ruling Netflix based on amount of Content



- **Overall Top Netflix Actor :**

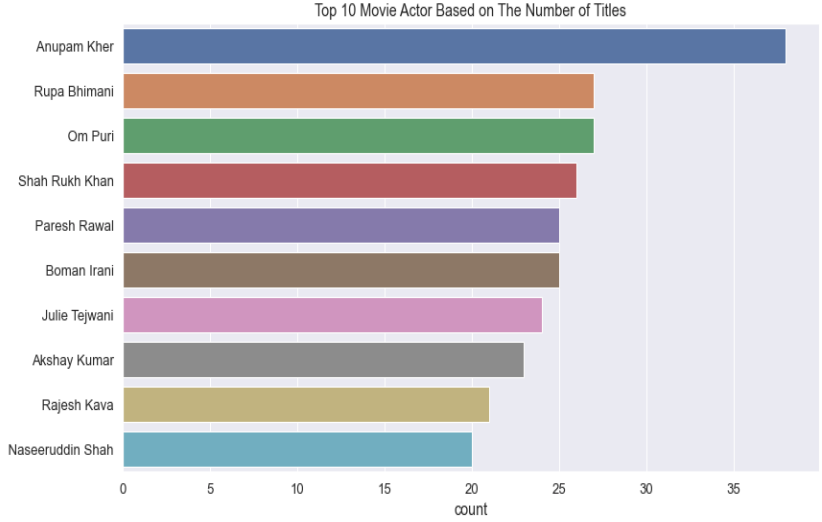
  Anupam Kher with 38 Titles

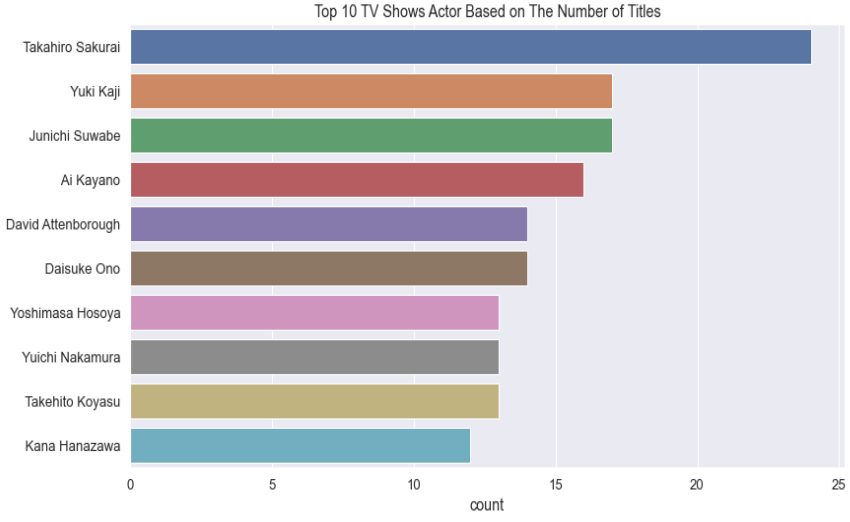
- **Top TV show Actor :**

  Takahiro Sakurai with 24 Titles

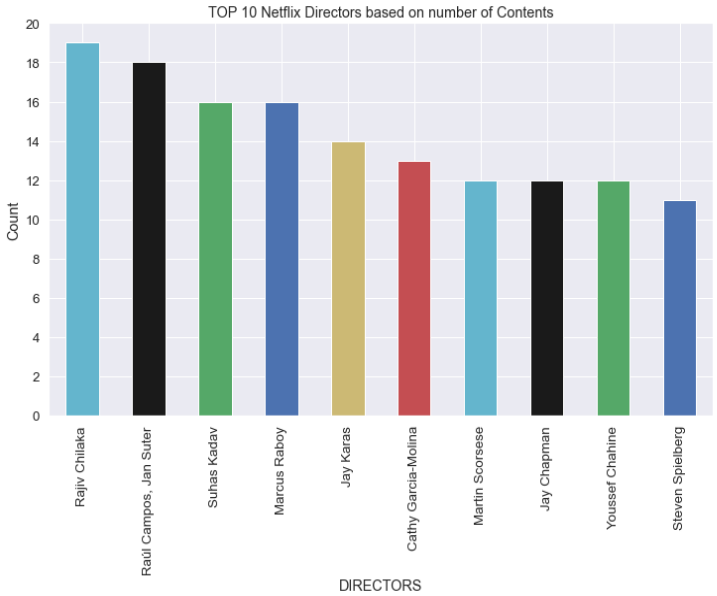
- **Top Netflix Movie Actor :**

  Anupam Kher with 38 Titles





Top Directors and Their Content



- **Overall Top Netflix Director :**

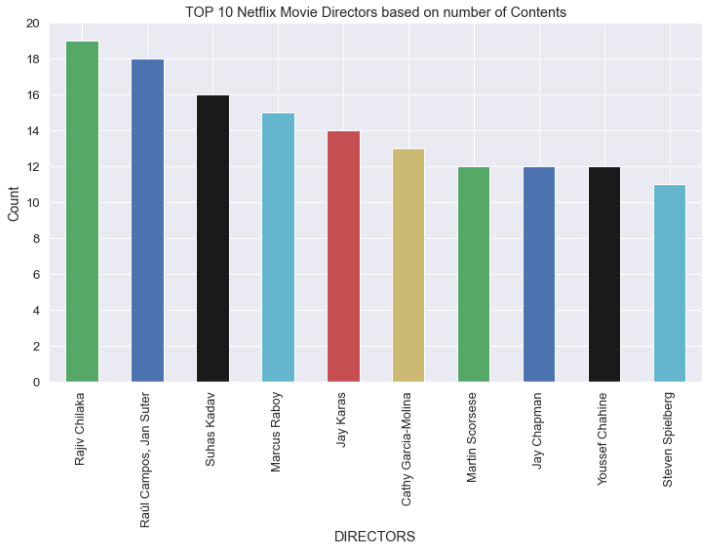
  Rajiv Chilaka with 19 Titles

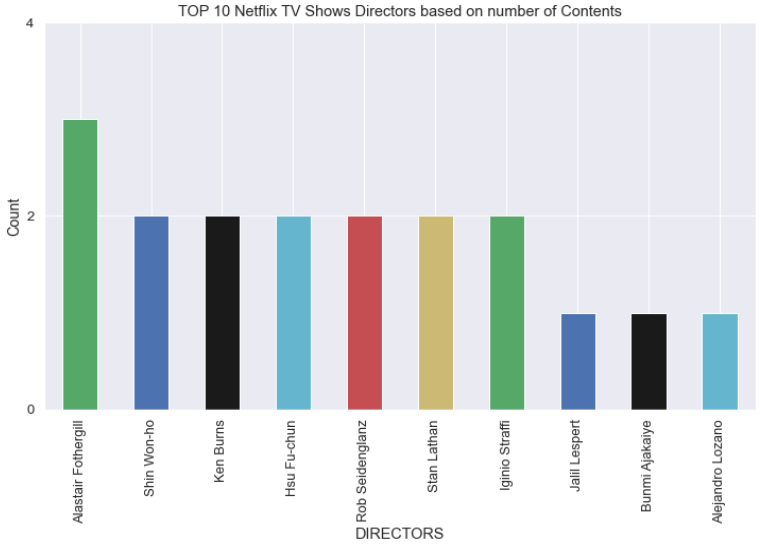
- **Top TV show Director :**

  Alastair Fothergill with 3 Titles

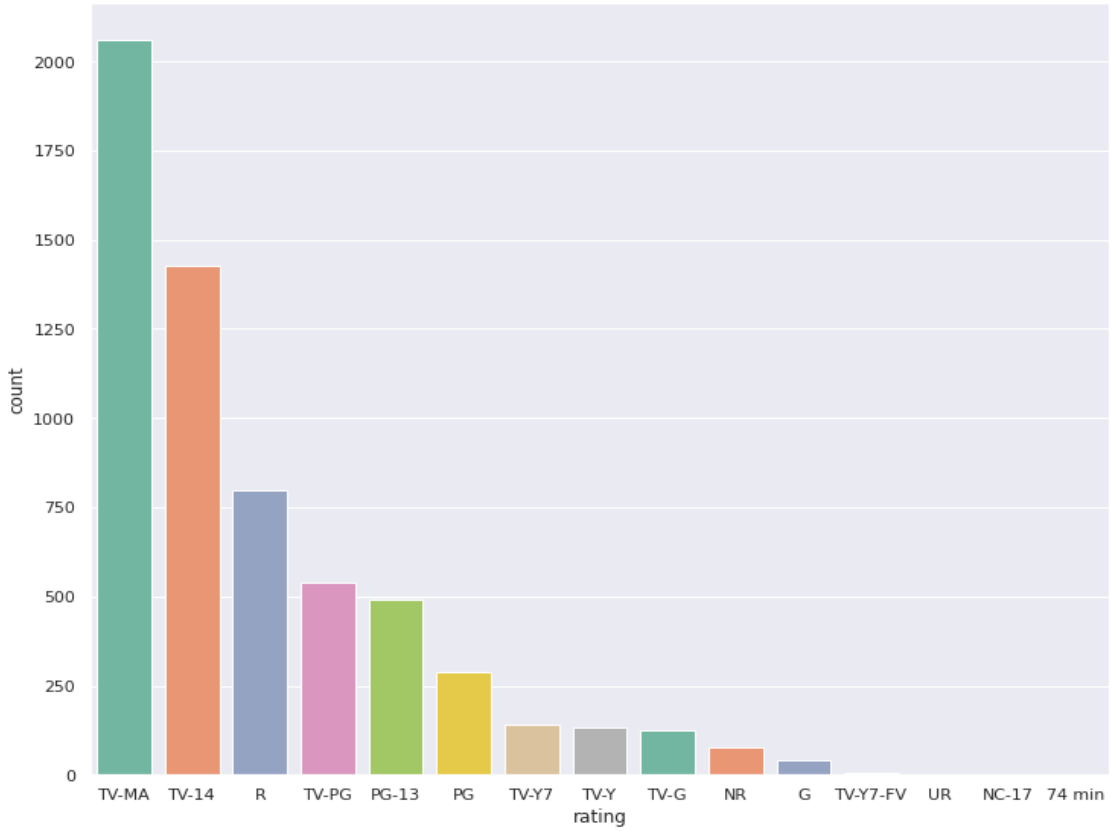
- **Top Netflix Movie Director :**

  Rajiv Chilaka with 19 Titles

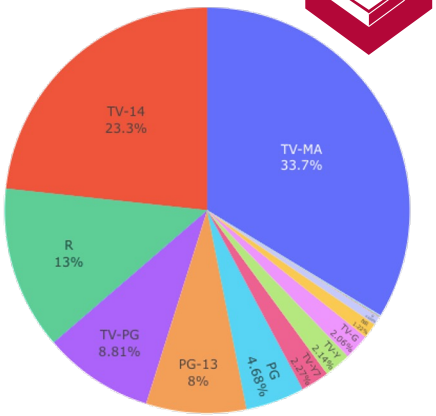




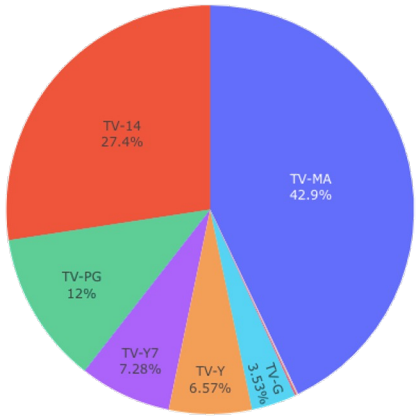
5. Rating and Content distribution on Netflix



Movie Rating

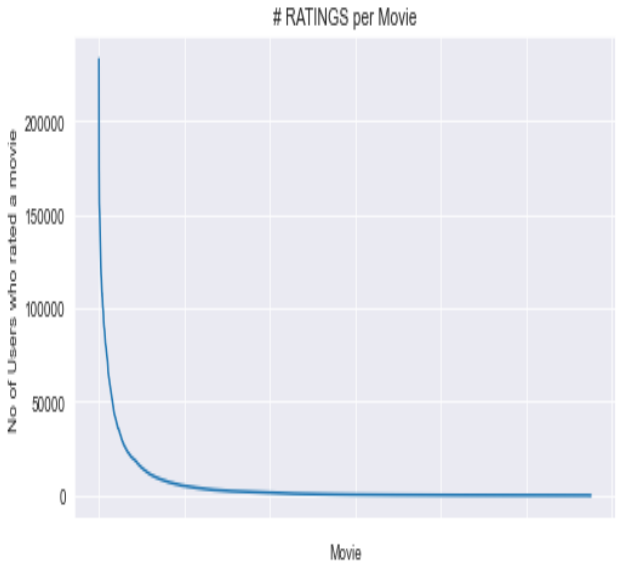


TV- Show Rating

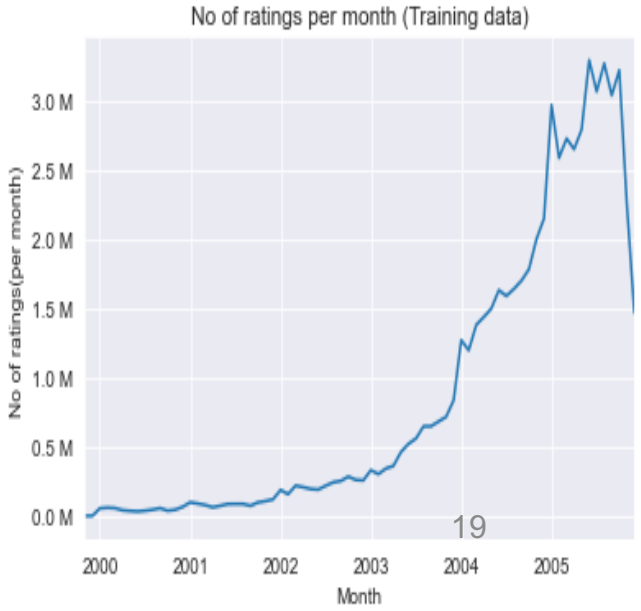


- The largest count of Netflix content is made with a “TV-MA ” rating. Which is content designed for mature audiences only
- The second highest content is “TV-14” contains material that parents or adult guardians may find unsuitable for children under the age of 14

6. User Rating Trends



- Data indicates the rating tends to be relatively on positive side which is 3 or more
- One probable reason is unsatisfied customers tend to just leave the platform instead of making efforts to rate assuming low rating movies mean they are generally bad
- The rating given by users are skewed (almost 90%)indicates some popular movies are rated by huge population



Recommender System Development

Content Based Recommender System

- The goal is to identify ‘similar’ movies to the user input.
- In this case, movie name will be input and the algorithm will predict movies with similar attributes (cast, director, country, rating and genres as features)
- Algorithm will generate a matrix of these attribute features to generate cosine-similarity matrix
- The list of movies will be reordered based on their similarity score

Advantages

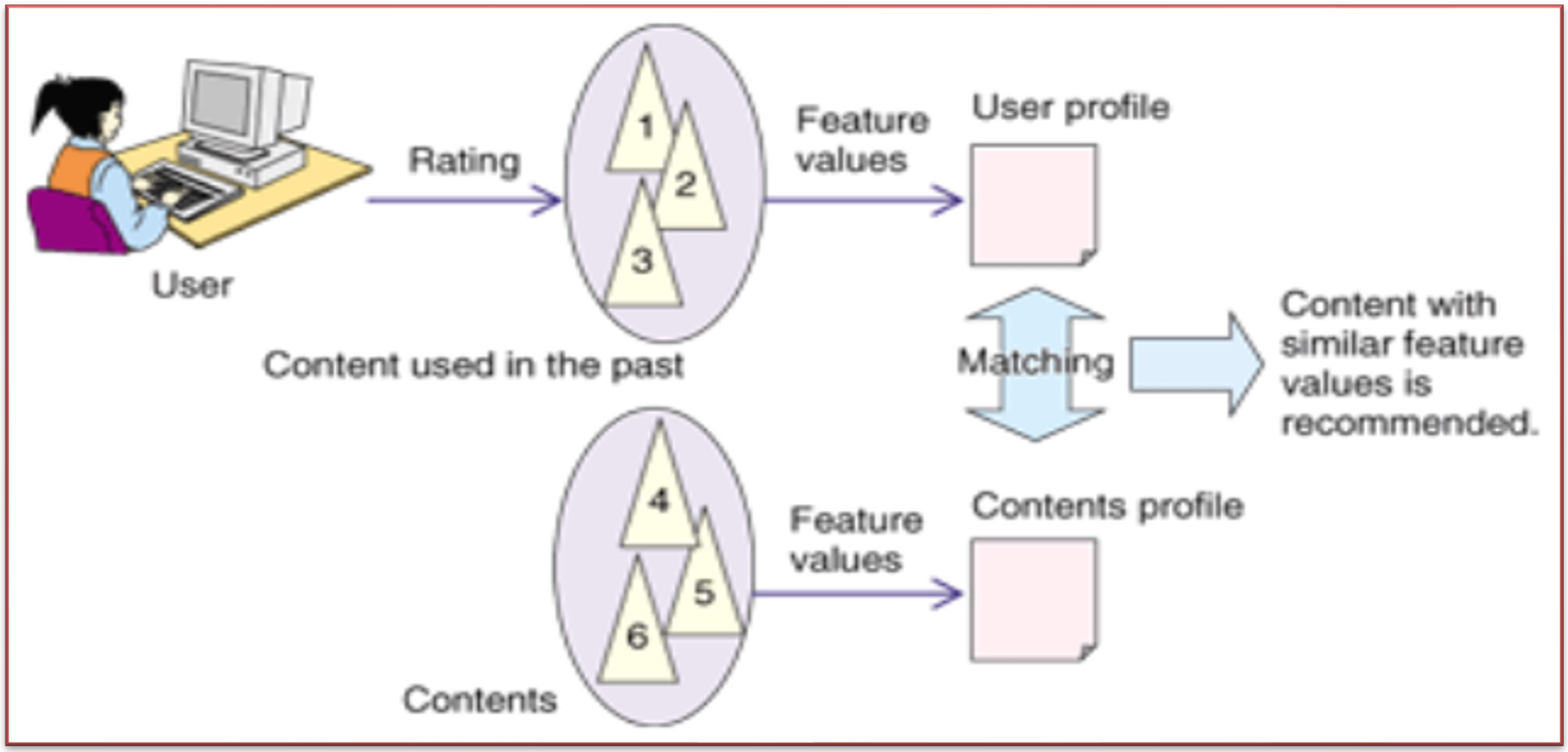
- User Independence as it takes only particular user’s preferences as input
- Transparency based on features
- New items can be suggested even before rating them
- Easily Scalable

Disadvantages

- Limited content analysis
- Over Specialization
- New user accuracy is low

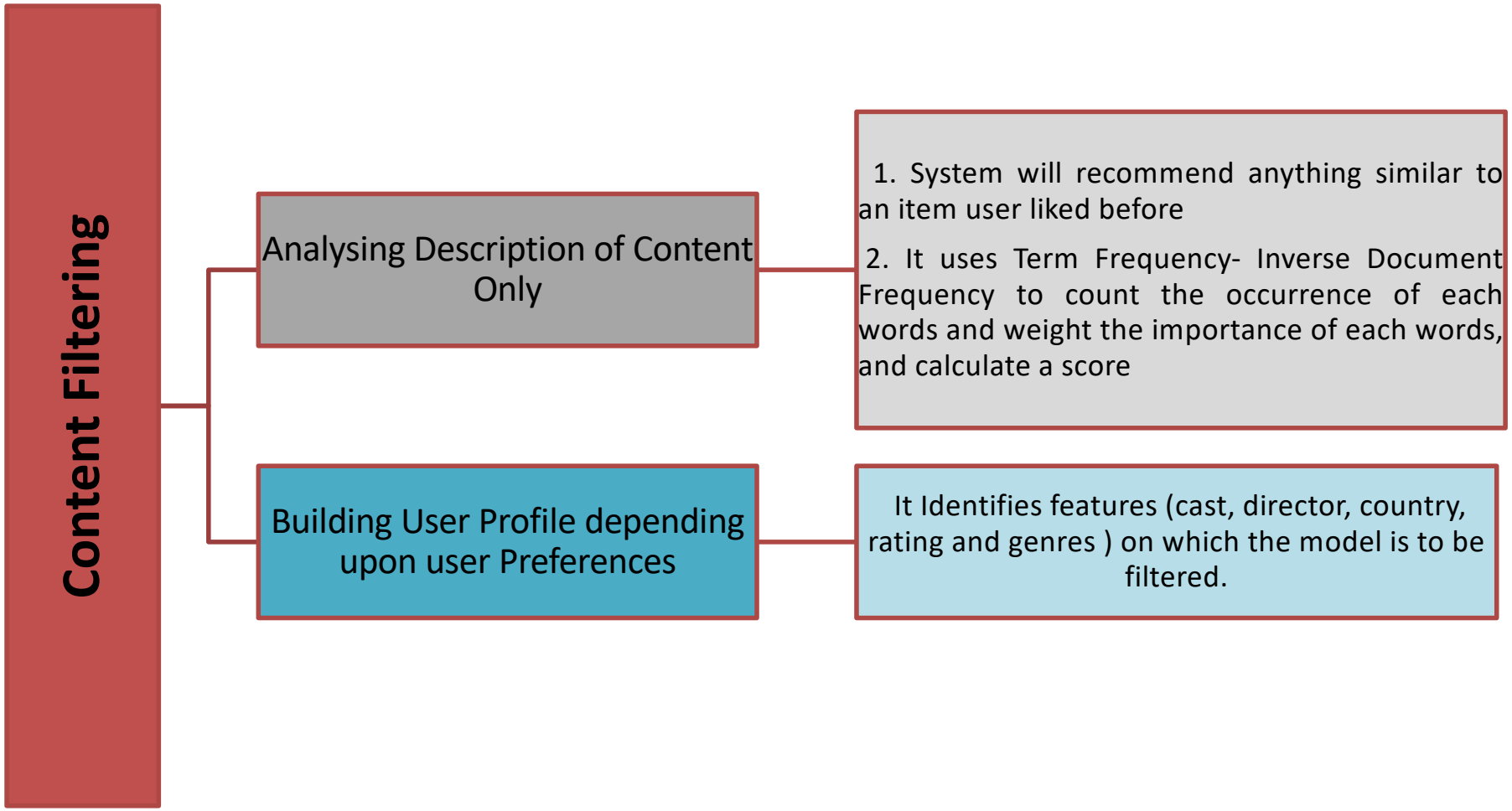
Ref: https://www.educative.io/edpresso/what-is-content-based-filtering , http://findoutyourfavorite.blogspot.com/2012/04/content-based-filtering.html
https://towardsdatascience.com/deep-dive-into-netflixs-recommender-system- 341806ae3b48

Model



Ref: http://findoutyourfavorite.blogspot.com/2012/04/content-based-filtering.html

Content Filtering Techniques



Ref: https://towardsdatascience.com/introduction-to-two-approaches-of-content-based-recommendation-system-fc797460c18c

Use of Cosine Similarity to identify Similar Users based on the Features

- To estimate similarity between two users/ movies using Euclidean distance between two points seems to be easiest way
- For multiple users/ movies, the angle between the lines joining the points to the origin can be used to identify pattern
- Increased angle between lines, indicates decreased similarity and vice versa
- If the angle is zero, then the users are very similar/identical
- The cosine of an angle is a function that decreases from 1 to -1 as the angle increases from 0 to 180
- Hence, we will use cosine of the angle to find the similarity between two users
- This method will help to bring all users to the same level by removing their biases

For both the approaches, the model has been implemented using Cosine Similarity

Ref : https://en.wikipedia.org/wiki/Cosine_similarity
https://realpython.com/build-recommendation-engine-collaborative-filtering/

Sample Model Output

Top 5 Movies Recommendations for movie PK

Analysing Description of Content Only	User Profile Depending Upon Preferences

3714	Unbroken
4221	Merku Thodarchi Malai
7129	Jhansi Ki Rani
906	Have You Ever Seen Fireflies? - Theatre Play
4306	ROMA

1114	3 Idiots
8391	The Legend of Michael Mishra
4790	Anthony Kaun Hai?
1022	Taare Zameen Par
4507	Sanju

Common Description words:

India, Amir Khan, International, Political Corruption, Comedies, country, independence

Common Director, Cast, Genre:

- Director = Rajkumar Hirani
- Actor = Amir Khan/ Anushka Sharma
- Genre = Comedy

Highlights

Approach	Analysing Description of Content Only	User Profile Depending Upon Preferences
Advantages	• **New item** problem can be solved if sufficient descriptions available • **Transparent system:** same approach for all recommendations	• **User Independence**: It only analyses item and user profile for recommendation • **No Cold Start** : new items can be suggested before being rated by a substantial number of users
Disadvantages	• **Over-specialization**: It might recommend items similar to those that already watched (No cross content)	• **Limited Content Analysis**: Item recommendation based only on user preferences (No surprise factor) • **New user**: lack of profile information can lead to incorrect recommendations

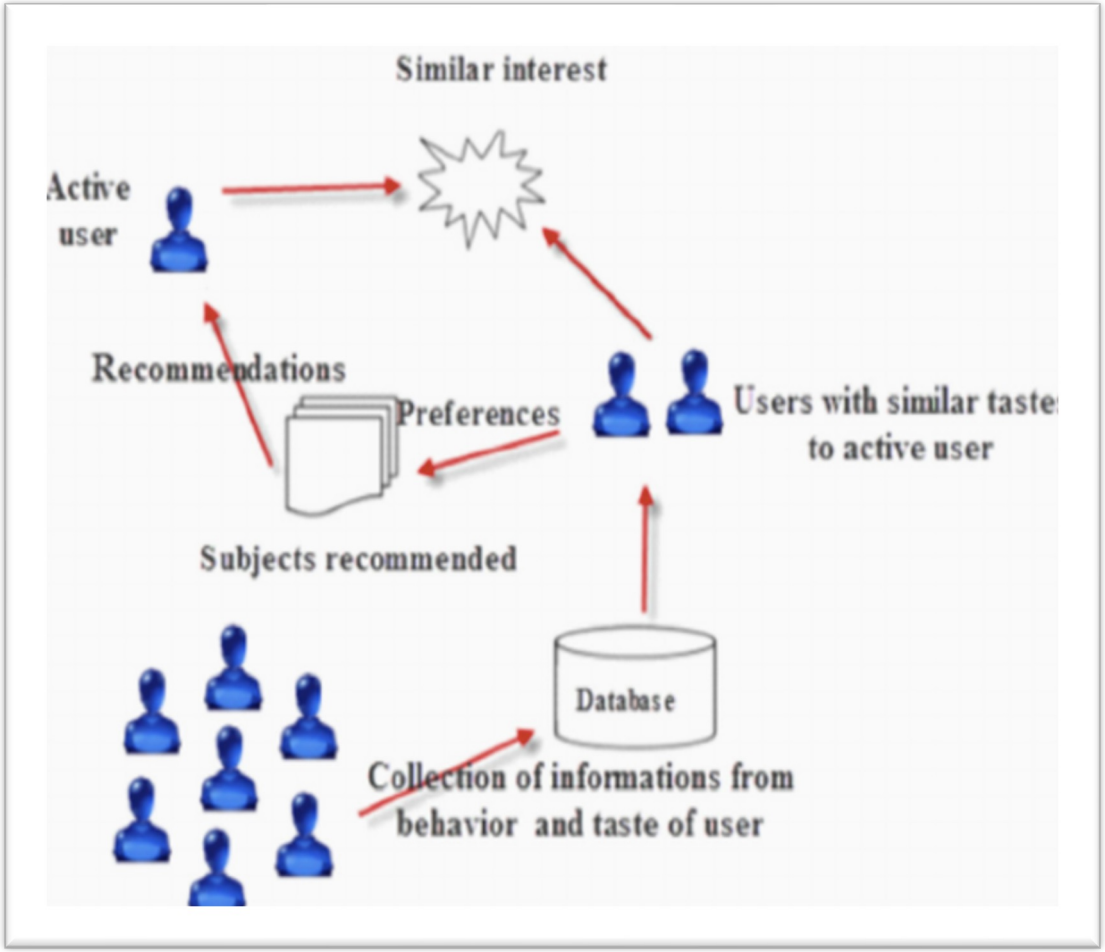
Collaborative Recommender System

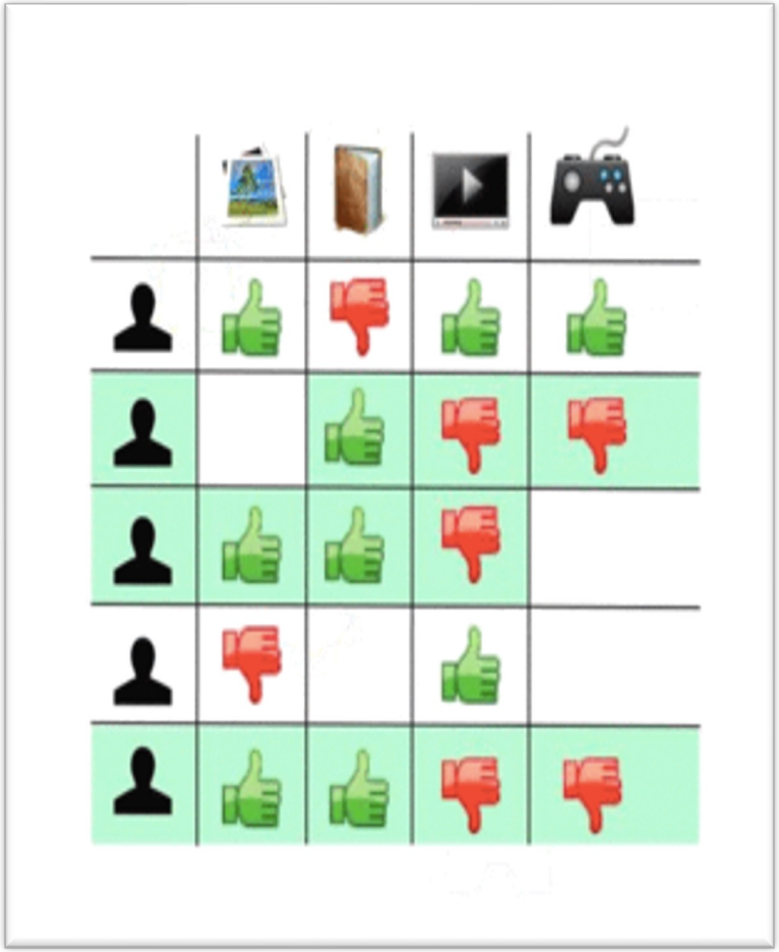
- The dataset will be focused on a **set of items** (Movies) and a **set of users** who have **reacted** (ratings) to some of the items.
- Steps Involved:
  1. Find similar users or movies based on the ratings.
  2. Predict the ratings of the movies that are not yet rated by a user
  3. Measure the accuracy of your predictions to find best fit model
  4. Build The Recommender System based on chosen model

Advantages	Disadvantages
• No domain knowledge is required • Model can help users discover new interests	• Cold Start problem • Hard to include side features

Ref: https://developers.google.com/machine-learning/recommendation/collaborative/summary
https://towardsdatascience.com/deep-dive-into-netflixs-recommender-system- 341806ae3b48
https://towardsdatascience.com/various-implementations-of-collaborative-filtering-100385c6dfe0

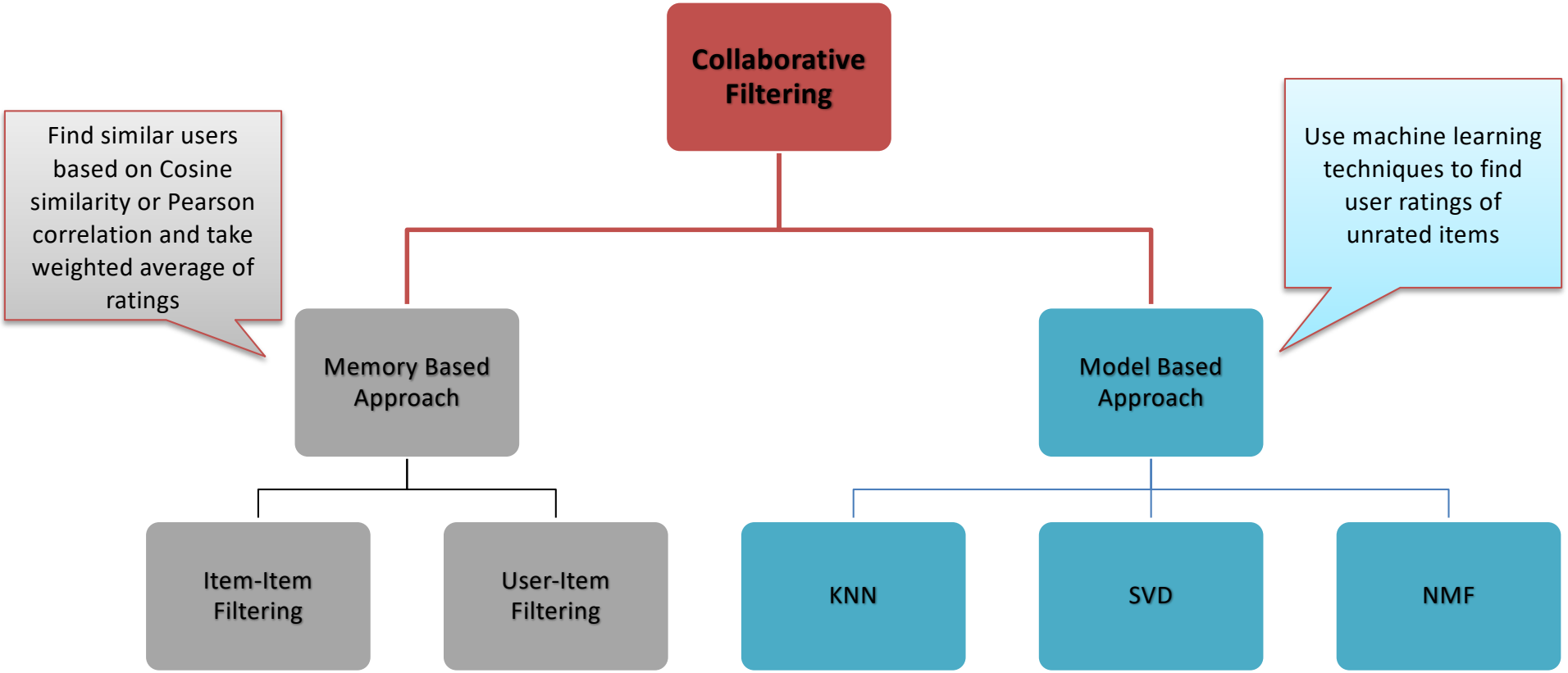
Model





Ref : https://www.researchgate.net/figure/Simple-description-work-of-collaborative-filtering_fig1_273268530

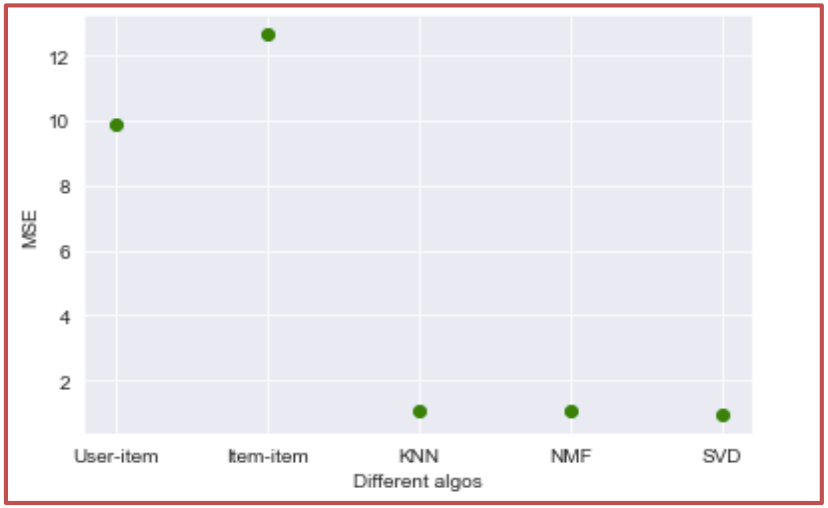
Collaborative Filtering Techniques



Ref: https://developers.google.com/machine-learning/recommendation/collaborative/summary
https://lazyprogrammer.me/tutorial-on-collaborative-filtering-and-matrix-factorization-in-python/
https://towardsdatascience.com/various-implementations-of-collaborative-filtering-100385c6dfe0
http://surpriselib.com/

Model Results

Model Comparison



The plot shows of MSE obtained from different approaches.

Technique	User-Item	Item-Item	KNN	NMF	SVD
MSE	9.9	12.65	1.07	1.06	0.95

- **The SVD algorithm gives Best Results.**

Sample Model Output

For User with ID:30878 top 5 recommended movies using SVD algorithm are:

1. Isle of Man TT 2004 Review
2. Character
3. Sick
4. Paula Abdul's Get Up & Dance
5. What the #$*! Do We Know!?

The user's (30878) average rating for movies are 3.5

Ref: http://surpriselib.com/
https://lazyprogrammer.me/tutorial-on-collaborative-filtering-and-matrix-factorization-in-python/

Highlights

Technique	Memory Based Approach	Model Based Approach
Advantages	• Easy Creation and Explain- ability of results	• Dimension reduction deals with missing/sparse data
Disadvantages	• Non-Scalable: Performance reduces when data is sparse • May require lots of memory and CPU time for large scale dataset • Cold Start Problem	• Interference is intractable because of hidden /latent factors • Increased complexity may lead to expensive implementation • Cold Start Problem

Ref: https://en.wikipedia.org/wiki/Collaborative_filtering

Evaluation

Lessons Learned :

Adopt Hybrid Model Based on the Case Scenario

If user is new to Netflix and knows his preferences

Use Content Based Recommender Model

Build User Profile based on User Preferences and provide appropriate Recommendations

If new movie/ TV show added to the Netflix

Use Content Based Recommender Model

Add Movie/ TV show with detailed information and analyze content while providing recommendations

Lessons Learned :

Adopt Hybrid Model Based on the Case Scenario

Tenured user has preference patterns

Use Collaborative Filter Recommender Model

Build User-Item Model : Create user matrix on rated/viewed/ liked/ clicked items and recommend similar items which user is unaware of

Dataset has detailed Information about movies and ratings

Use Collaborative Filter Recommender Model

Build Item-Item Model: Based on movie ratings similarity, find the common users to recommend new movies

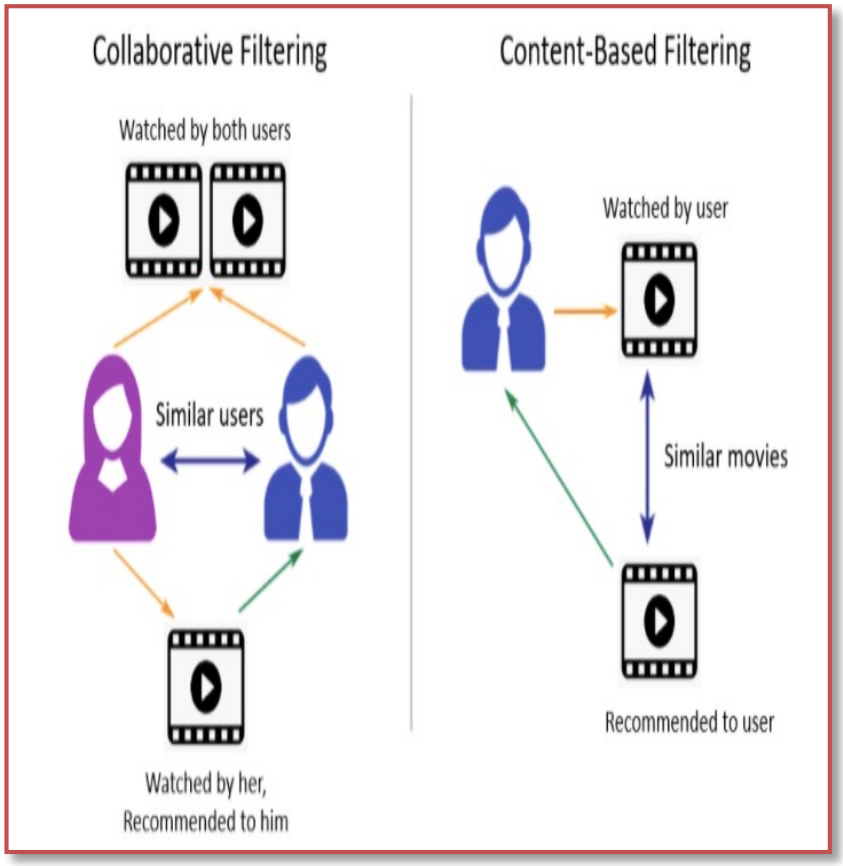
Predict the ratings of the movies that are not yet rated by a user

Use Collaborative Filter Recommender Model

Use machine learning techniques to find user ratings of unrated items

Conclusion

- Content based model is best suited for new user or item as it doesn't consider other user preferences and patterns. However, it may lead to limited recommendation options and nosurprise factor to the user
- Collaborative model is best suited to predicted ratings of item that has not been watched by user. However, it may face Cold Start Problem : If we do not have sufficient data about users (0.15%) and movies (19.5%) we are unable to form any relation between them
- With use of hybrid model, we can try to overcome limitations of both the approaches to maximize customer satisfaction and retention
- Best fit recommender model could be a weighted hybrid based on the use case scenario:
  - New vs Tenured User/ Item
  - Availability of data points



Ref: https://www.kdnuggets.com/2019/11/content-based-recommender-using-natural-language-processing-nlp.html

Data Source

Dataset Available At:

- For content based Model:
  https://www.kaggle.com/shivamb/netflix-shows

- For collaborative filtering based Model:
  https://www.kaggle.com/netflix-inc/netflix-prize-data

Thank You