Netflix analysis and recommendation system

PES UNIVERSITY
EC CAMPUS

V P Srinidhi
PES2UG19CS439

Prerana Umakant Bandekar
PES2UG19CS297

M Satvika
PES2UG19CS207

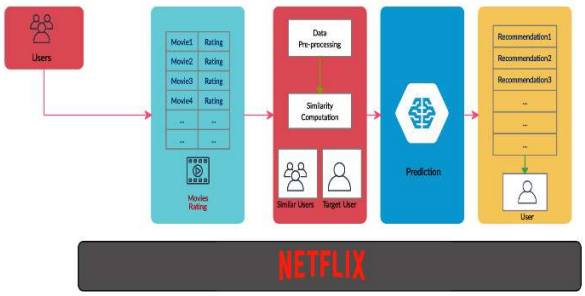
Rishitha Chowdary M
PES2UG19CS205

Abstract* – Netflix is a media service provider that is based out of America. It provides movie streaming through a subscription model. It includes television shows and in house produced content along with movies. In this paper we have predicted recommendation based on content, director and genre using machine learning algorithm (recommendation system). The dataset shows *listings of all the movies and tv shows available on Netflix, along with details such as - cast, directors, ratings, release year, duration, etc.

Keywords – movies, recommendation system, ml models, ratings, director.

I. INTRODUCTION

Netflix has taken up an active role in producing movies and TV shows. The company is heavily data driven. Netflix lies in the middle of the internet and storytelling. They are inventing new internet television. Their main source of income comes from users' subscription fees. They allow users to stream data from a wide range of their movies and TV shows at any time on a variety of internet-connected services. The primary asset of Netflix is their technology. Especially their recommendation system. Information filtering systems deal with removing unnecessary information from the data stream before it reaches a human. Recommendation systems deal with recommending a product or assigning a rating to item. They are mostly used to generate playlists for the audience



Netflix is all about recommending the next content to its user. The only question they would like to answer is 'How to personalize Netflix as much as possible to a user?'. Though it is a single question, it is almost everything Netflix aims to solve. Recommendation is embedded in every part of their site.

Recommendation starts when you log into Netflix. For example, the first screen you see after you log in consists of 10 rows of titles that you are most likely to watch next. Awareness is another important part of their personalization. They let their audience know how they are adapting to their tastes.

The objective of this project is to understand more about recommendation system using Machine Learning algorithms. We have used Netflix Movies and TV Shows dataset from Kaggle. The dataset contains the movies and tv shows of different languages across the world. It has 8,807 rows and 12 observations including both predictors and dependent variable.

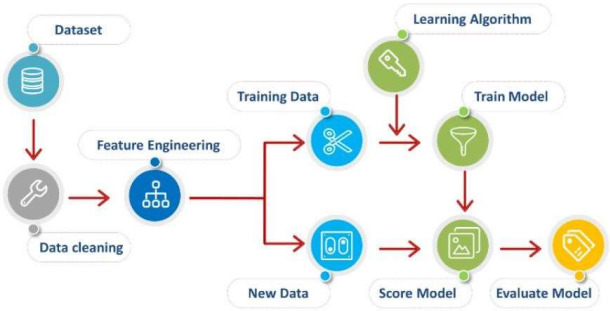
II. PROBLEM STATEMENT

Project's objective is to create a recommendation system based on different observations like director, genre. We'll use machine learning algorithm to achieve this goal.

III. METHODOLOGY

The following steps are implemented to build a required machine learning model to create a :

A. Data Loading from the CSV file.
B. Understanding the data.
C. Visualizing the data.
D. Preparing the data for the model.
E. Modeling.
F. Model Evaluation.



A. Data Loading from the CSV file

Initially, all the basic libraries numpy, pandas, matplotlib, seaborn, sklearn are imported. There are many other machine learning libraries which are imported accordingly.

The dataset contains movies and tv shows of different languages across the world. The CSV file loaded into Pandas data-frame using read_csv() function.

```
In [3]: import numpy as np
        import pandas as pd
        df = pd.read_csv(r'netflix_titles.csv')
        df.head()
```

Out[3]:

	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	India	September 24, 2021	2021	TV-MA	2 Seasons	International TV Shows, Romantic	In a city of coaching centers known

```
In [4]: df.shape
```

Out[4]: (8807, 12)

```
In [5]: df.isnull().sum()
```

Out[5]:
```
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
```

```
In [6]: df = df[df['director'].notnull()]
In [7]: df.shape
```

Out[7]: (6173, 12)

```
In [8]: df[df['rating'].isnull()]
```

Out[8]:

show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description

B. Understanding the data

Dataset has 8,807 rows and 12 observations including both predictors and dependent variable. This Dataset contains a combination of numerical and categorical data.

The following observations are made after seeing the data in the data-frame:

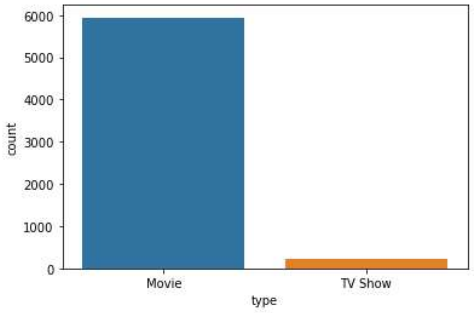
1. Categorical features are show_id ,type ,title ,director cast, country.
2. Numerical features are date_added, release_year rating ,duration.

C. Visualizing the data

The dataset is visualized by plotting few graphs/plots using matplotlip and seaborn libraries.

Graph to compare the number of movies to tv shows&documetories.

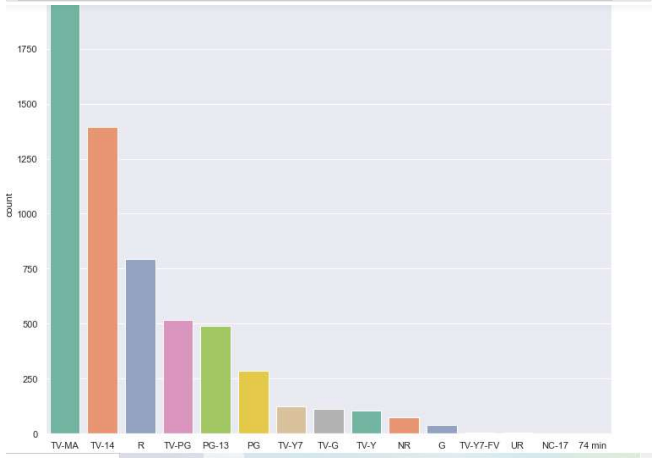
```
import seaborn as sns
ax = sns.countplot(x="type", data=df)
```



By seeing the plot, we could conclude that there are a lot of movies in Netflix when compared to tv shows.so to help the audience to pick a movie or a show of there liking a recommendation system is important.

Graph about Movies with different ratings

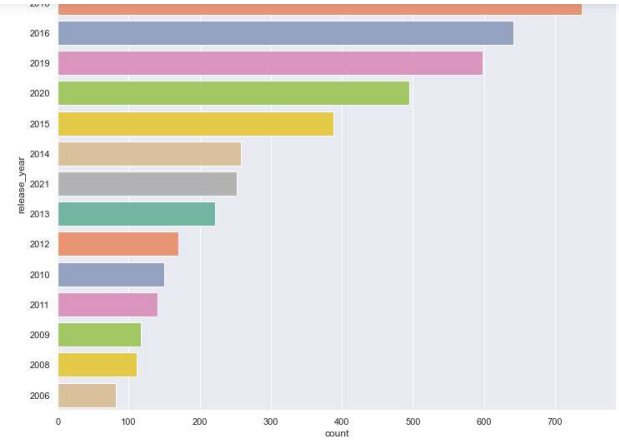
```
import matplotlib.pyplot as plt
plt.figure(figsize=(12,10))
sns.set(style="darkgrid")
ax = sns.countplot(x="rating", data=netflix_movies, palette="Set2", order=netflix_movies['rating'].value_counts().index[0:
```



By visualizing the above bar plot, we have come to a conclusion, that there are a lot of movies which can be suggested to audience based on ratings.

Graph about movies and their release date

```
plt.figure(figsize=(12,10))
sns.set(style='darkgrid')
ax = sns.countplot(y='release_year', data=netflix_movies, palette='Set2', order=netflix_movies['release_year'].value_count
```



By seeing the above plot , we can develop a recommendation system based on release date and can sort them according to their release date.

D. Preparing the data for model

Initially, all the basic libraries numpy, sklearn are imported. There are many other machine learning libraries which are imported accordingly

```
import pandas as pd
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.metrics.pairwise import cosine_similarity
from sklearn.feature_extraction.text import CountVectorizer
```

E. Modeling

We have used Cosine similarity and bag of words to recommend movies based on the input data.

```
In [7]: # First model is based on the description of a particular movie using cosine similarity and bag of words.
        # word vector will be calculated using TF-IDF

        tfidf = TfidfVectorizer(stop_words='english')
        df['description'].fillna('')
        word_matrix = tfidf.fit_transform(df['description'])

In [8]: print(word_matrix.shape)

        (5700, 14767)

In [9]: cs = cosine_similarity(word_matrix, word_matrix)

In [10]: top_7_recommendations = content_based_recommendation("Gunglands",cs)

In [11]: print("Top recommendations based on the plot:")
         print(top_7_recommendations)

         Top recommendations based on the plot:
         4364    My Little Pony Friendship Is Magic: Best Gift ...
         4588                                        My Friend Pinto
         7110                        Jack and the Cuckoo-Clock Heart
         5485                                              Ram Jaane
         555                                             Snowpiercer
         6641                                            Dragonheart
         2314                                               Stardust
         Name: title, dtype: object
```

F. Model Evaluation

After performing tuning of parameter, we got some pretty good test accuracy of about 90%.

a) Model 1: Random Recommendation

This is done based on the genre

Given a movie, we take all the movies of that genre and randomly display 7 to the user

```
Top recommendations based on the plot:
4364    My Little Pony Friendship Is Magic: Best Gift ...
4588                                        My Friend Pinto
7110                        Jack and the Cuckoo-Clock Heart
5485                                              Ram Jaane
555                                             Snowpiercer
6641                                            Dragonheart
2314                                               Stardust
Name: title, dtype: object
```

Model 2: Content based model using plot

We take the plot, use tf-idf measure the find the vocabulary

Then we use cosine similarity and output 7 most similar movies

```
Top recommendations based on the genre:
11      Bangkok Breaking
1223              Dealer
3356         Nowhere Man
Name: title, dtype: object
```

Model 3: Content based model using director, genre, and cast

We process the data to remove spaces and make it lowercase

Create a word soup

Then again use cosine similarity between them and output 7 most similar movies

```
Top Movies based on combination of Cast, director and Genre:
6433            Cats & Dogs: The Revenge of Kitty Galore
3016                                                Hop
3248                         The Knight Before Christmas
1681                 The Princess Switch: Switched Again
2858    Calico Critters: Everyone's Big Dream Flying i...
2188                                         Sugar High
1304          Animals on the Loose: A You vs. Wild Movie
Name: title, dtype: object
```

IV. Making Prediction on New Inputs

We have hardcoded inputs for making predictions. We took the movie 'ganglands' for all the models and tested what were the recommendations received

Model 1:

```
In [10]: top_7_recommendations = content_based_recommendation("Ganglands",cs)

In [11]: print("Top recommendations based on the plot:")
         print(top_7_recommendations)

Top recommendations based on the plot:
4364    My Little Pony Friendship Is Magic: Best Gift ...
4588                                    My Friend Pinto
7110                    Jack and the Cuckoo-Clock Heart
5485                                      Ram Jaane
555                                       Snowpiercer
6641                                      Dragonheart
2314                                         Stardust
Name: title, dtype: object
```

Model 2:

Model 3:

```
In [18]: top_7_director = recommendation_director("Ganglands")

In [22]: print("Top Movies based on combination of Cast, director and Genre:")
         print(top_7_director)

Top Movies based on combination of Cast, director and Genre:
6433              Cats & Dogs: The Revenge of Kitty Galore
3016                                                  Hop
3248                          The Knight Before Christmas
1681                  The Princess Switch: Switched Again
2858    Calico Critters: Everyone's Big Dream Flying i...
2188                                          Sugar High
1304           Animals on the Loose: A You vs. Wild Movie
Name: title, dtype: object
```

V. Conclusion

From the project we have learned the working of machine learning algorithm on our dataset. Using of Recommendation systems turn out to be a great asset as they show increase in the revenue and sales and there's also a growth in the user satisfaction rate. The better our recommendation system is the better are the results in terms of business growth. Users always prefer applications which will ease their work of searching and finding what they want by showing them recommendations of what they might like based on their personal interests. The Netflix recommendation system uses exactly this technique and thus is the most used video streaming platform across the globe. The model we built is put into action when integrated with websites.

References

[1] Introduction to Machine Learning with Python by Andreas C. Muller and Sarah Guido.

[2] https://machinelearningmastery.com/hyperparameters-forclassification-machine-learning-algorithms

[3] https://ieeexplore.ieee.org/abstract/document/8993091/auth

[4] https://towardsdatascience.com/recommender-systems-in-practice-cef9033bb23a

[5] https://developers.google.com/machine-learning/recommendation/content-based/basics