ALBUKHARY INTERNATIONAL UNIVERSITY

NAME AND ID:	COURSE NAME AND CODE :	ASSIGNMENT TITLE:
SIDY YAYA TOURE	DATA MINING	NETFLIX MOVIE ANALYSIS
AIU22102009	CCS2313	REPORT

Report of Assignment:

---

1) Project Background:

In the modern digital world, the entertainment industry has experienced a revolution, as more online streaming media have been developed. Such platforms as Netflix, Amazon Prime, or Disney+ provide users with thousands of movies and TV shows, which is overwhelming and it is challenging to select one among the many options. On the one hand, this richness gives more choices, but on the other hand, it is hard to find content within a very short time that suits the personal preferences of users.

To counter this challenge, recommendation systems have come up as a solution that is automated to propose content that is relevant to the user interests. These systems consider the trends in user behavior, including the history viewed, the ratings, likes and the search topics to determine what the user might find enjoyable to watch next. Such recommendation algorithms have significant contribution to the level of user engagement and retention by companies such as Netflix.

The Metflix Movie Recommendation System is the proposed project which is aimed at designing and developing a recommendation engine that will increase user experience by being able to give personal recommendations on the movies. The system will use the methods of machine learning and data analysis to suggest movies depending on their user preference and behavior. Various recommendation techniques like: content-based filtering,

collaborative filtering and hybrid techniques can be used to achieve accuracy and variety of recommendations.

Through the implementation of this project, users will not have to spend most of their time searching but watching movies of their choice. Moreover, this project shows how the concepts of artificial intelligence and data science can be applied in real-world situations, specifically, how these two aspects can be used to enhance the quality of decision-making and satisfaction of users in digital entertainment platforms.

1.1)Problem Statement and Issue:

The streaming sites such as Netflix have altered the manner in which people consume movies, though they have a downside; an excess number of options. The movies on the same platform may number to thousands, and rather than making the process easier, there is always a risk of scrolling through and even taking minutes or hours before settling on what particular movie to watch. This causes frustration and decision paralysis, where individuals opt to give up, watch the same movies or just have to accept something they are not even a fan of.

The point is that without the intelligent approach to the content recommendation, users are not capable of locating films that do suit their peculiar preferences. Such generic categories as Top Picks or Trending are not sufficient, as popularity may not be what certain user desires. Consequently, streaming services are at risk of losing the interest of the users and dissatisfaction.

Thus, the issue that this project is supposed to solve is obvious: customers require a privateized movie recommendation system that will be able to learn their taste and watching habits and suggest the movies that such customers are more likely to enjoy. In solving this, the project will alleviate the excessive sense of choice, save the users on time, and enhance the viewing experience.

1.2)The Assignment Objectives and Motivation:

The primary goal of the given project is to design a system of movie recommendation that will give personal recommendations according to the preferences and watch history of the users. Through recommendation methods like collaborative filtering, the system would assist the user in finding movies that suit them within a short period of time.

The idea of this project is driven by the fact that many streaming users have a real problem of having too many options that cause them to spend their time in a waste and get fatigued due to the decision. A good recommendation system would help not only in saving time but also the user satisfaction since better viewing experience will make it more exciting and entertaining.

2)Literature Review:

a) Recommendation systems have become an essential part of digital platforms, especially in entertainment and e-commerce, where the amount of available content is overwhelming. They provide personalized suggestions by analyzing user behavior, preferences, and patterns. Different approaches have been proposed and widely studied, with three of the most common being **content-based filtering**, **collaborative filtering**, and **hybrid methods**. **Lops, P., de Gemmis, M., & Semeraro, G. — "Content-based Recommender Systems: State of the Art and Trends"**

Springer chapter in *Recommender Systems Handbook*. link.springer.com+1

https://link.springer.com/chapter/10.1007/978-0-387-85820-3_3 link.springer.com

b) **Collaborative filtering** recommends items by finding patterns from the preferences of similar users. According to Su and Khoshgoftaar (2009), this method leverages the collective behavior of users to make predictions and is one of the most successful approaches in large-scale platforms like Netflix. However, collaborative filtering often struggles with the "cold start" problem when new users or items have insufficient data. **Su, X., & Khoshgoftaar, T. M. — "A Survey of Collaborative Filtering Techniques" (2009)**

Open access article: Advances in Artificial Intelligence, 2009. onlinelibrary.wiley.com

https://onlinelibrary.wiley.com/doi/10.1155/2009/421425

c) Collaborative filtering (CF) is one of the cornerstone approaches in recommender systems: instead of reasoning about movie content, it looks at the patterns in what *users* do and uses those patterns to predict what other users will like. Early surveys separate CF into memory-based (neighbourhood) methods and model-based methods; memory-based approaches are intuitive and easy to implement (e.g., "users like you also liked X"), while model-based approaches learn compact representations of users and items to make predictions more scalable and robust. This taxonomy and the main challenges of CF (data sparsity, cold start, scalability, and robustness to attacks) are well summarized in Su & Khoshgoftaar's comprehensive survey.

Su, X. & Khoshgoftaar, T. M. (2009). *A Survey of Collaborative Filtering Techniques.* onlinelibrary.wiley.com

3)METHODOLOGY:

The methodology of this project follows a structured approach to developing the Netflix movie recommendation system.

3.1)Project Framework

The project framework outlines the overall structure of how the Metflix recommendation system is designed and developed. It consists of **five main components**: data collection, preprocessing, feature extraction, recommendation engine, and evaluation.

3.2) Dataset:

The data set in this project is based on the Netflix movie dataset that is among the most popular benchmark datasets on the construction and testing of the recommendation systems. GroupLens Research maintains it and also has millions of ratings of movies given by actual users.

Context

This data is all about Movies That are available on Netflix Website **movies title, cast of the movie,desc of movies, duration, rating on IMDB, voted by people, year, genre, certificate**

source

This dataset comes from the IMDB website data is collected by using web scraping

```
# basics info
print("Shape:", df.shape)
print("Columns:", df.columns)
df.head()
```

```
Shape: (9957, 9)
Columns: Index(['title', 'year', 'certificate', 'duration', 'genre', 'rating',
       'description', 'stars', 'votes'],
      dtype='object')
```

3.3) Data Preparation:

a) **Data Collection :**

I dowmload this dataset from movie rating datasets, like the Netflixmovie recomendation dataset, that contains user-movie interactions, user-movie rating, and movie-metadata (genres, year of release, tags). This data is appropriate in constructing and testing the recommendation system.

b) **Load data:**

Firsly I load the dataset to start working on it.

c) **Data Preprocessing:**

The raw dataset is then cleaned and prepared to the high quality of the input to the model. This includes dealing with missing values, eliminating duplicates and coding categorical

attributes like genres and normalizing rating values. Interaction matrices of users and movies are created to indicate the preferences of the users on the various movies.

```
# Finding missing value
df.isnull().sum()
```

```
title            0
year           527
certificate   3453
duration      2036
genre           73
rating        1173
description      0
stars            0
votes         1173
dtype: int64
```

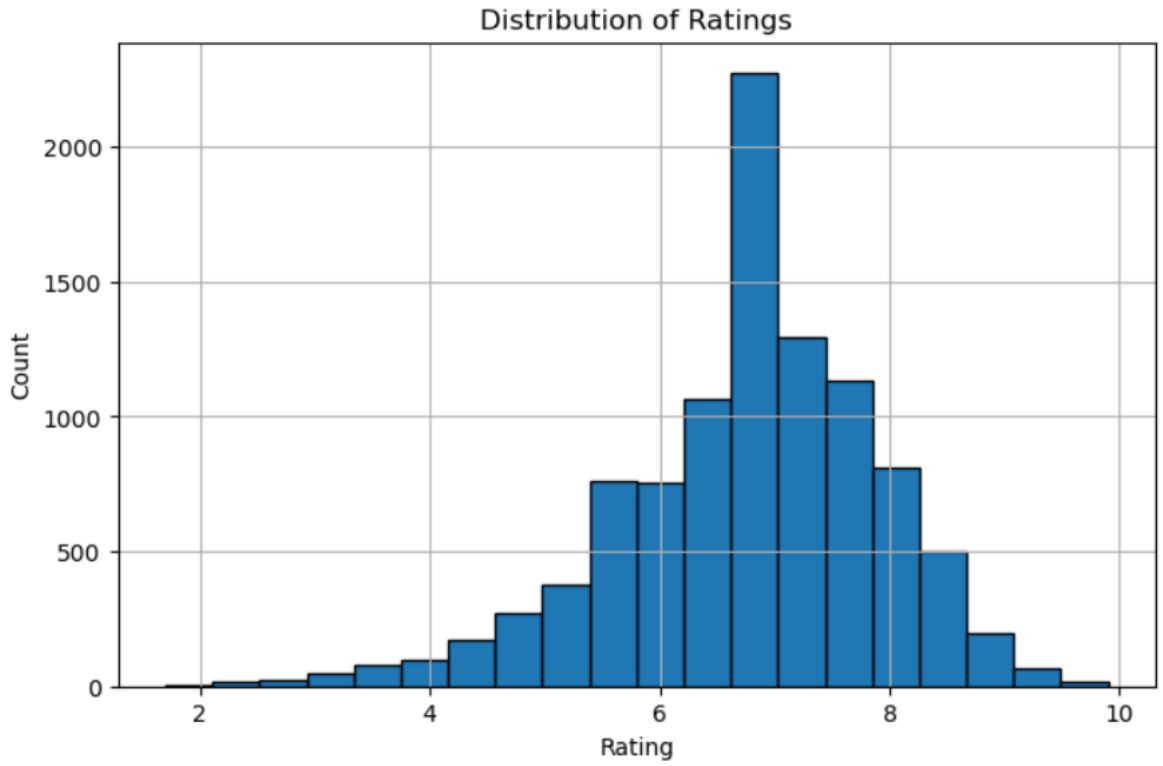
After cleanining missing value

```
df.isnull().sum()
```

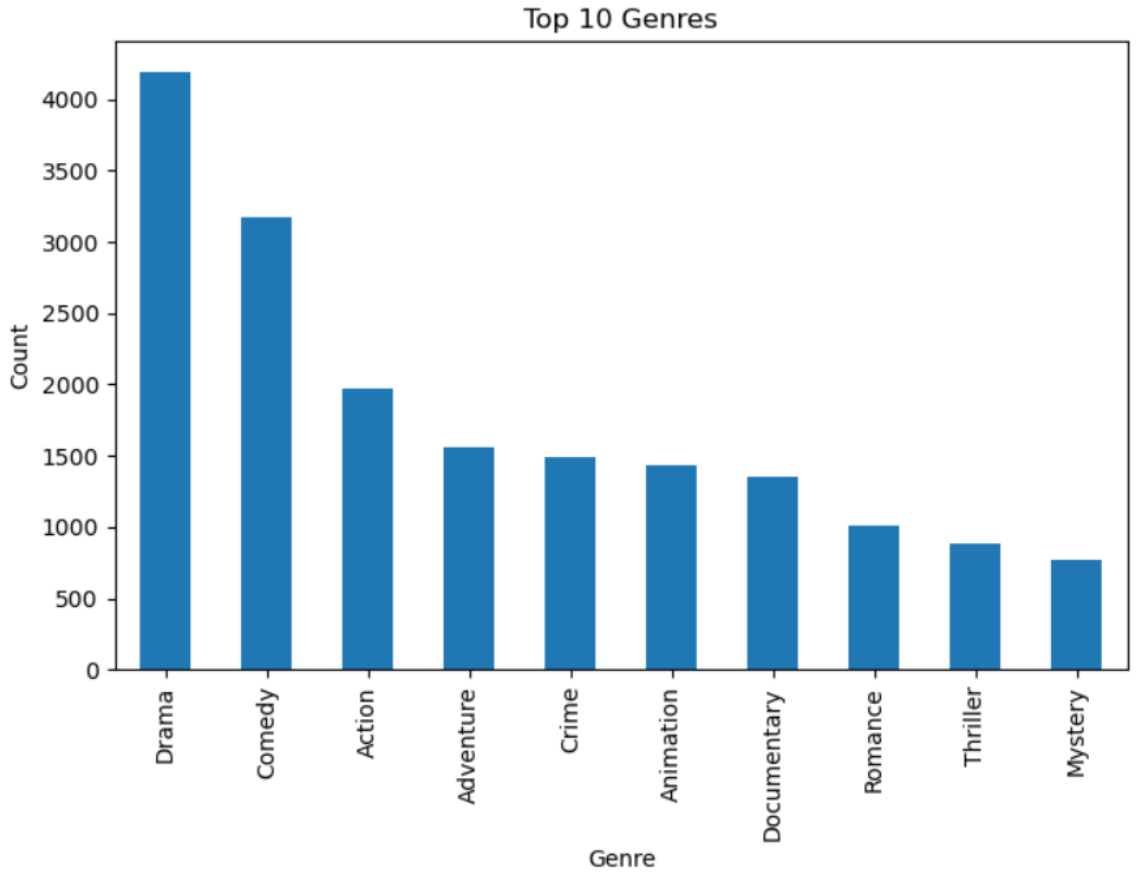
```
title          0
year           0
certificate    0
duration       0
genre          0
rating         0
description    0
stars          0
votes          0
dtype: int64
```

d) Exploratory data analysis:

Exploratory Data Analysis (EDA) is a crucial step in developing a movie recommendation system because it helps in understanding the structure, patterns, and characteristics of the dataset before building the model. In this project, EDA involves examining the distribution of user ratings to identify whether users tend to give higher or lower scores, analyzing the number of ratings each movie receives to distinguish between popular and less-watched films, and checking the activity level of users to understand which users contribute the most data.



Top 10 most viwed movies ploting



3.4) Approaches, algorithms:

In building the netflix movie recommendation system, three main approaches are considered: collaborative filtering, content-based filtering, and hybrid methods. But in this project we choose to work collaborative filtering.

4.0 Experiments and Analysis

Model Development or Recommendation Engine:
The recommendation engine is developed using one main approaches:

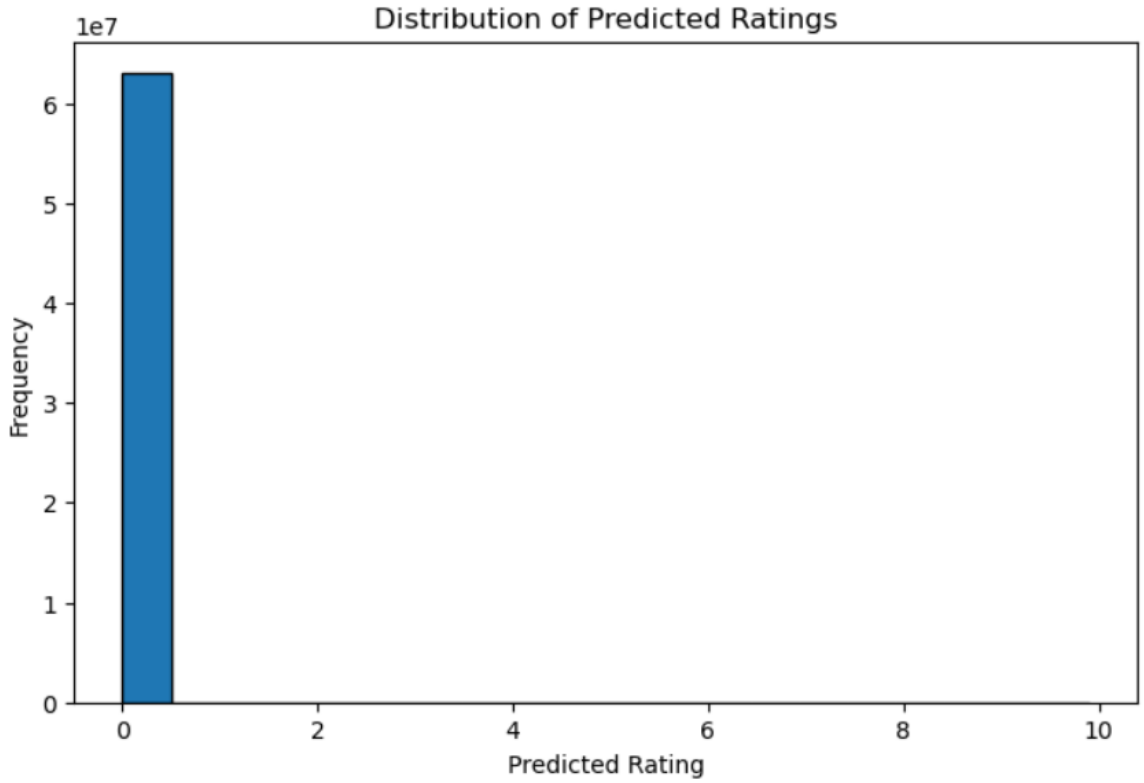
Collaborative Filtering (CF): Predicts user preferences based on the behavior and ratings of similar users or items. Both user-based and item-based collaborative filtering will be implemented.

Evaluation and Testing :

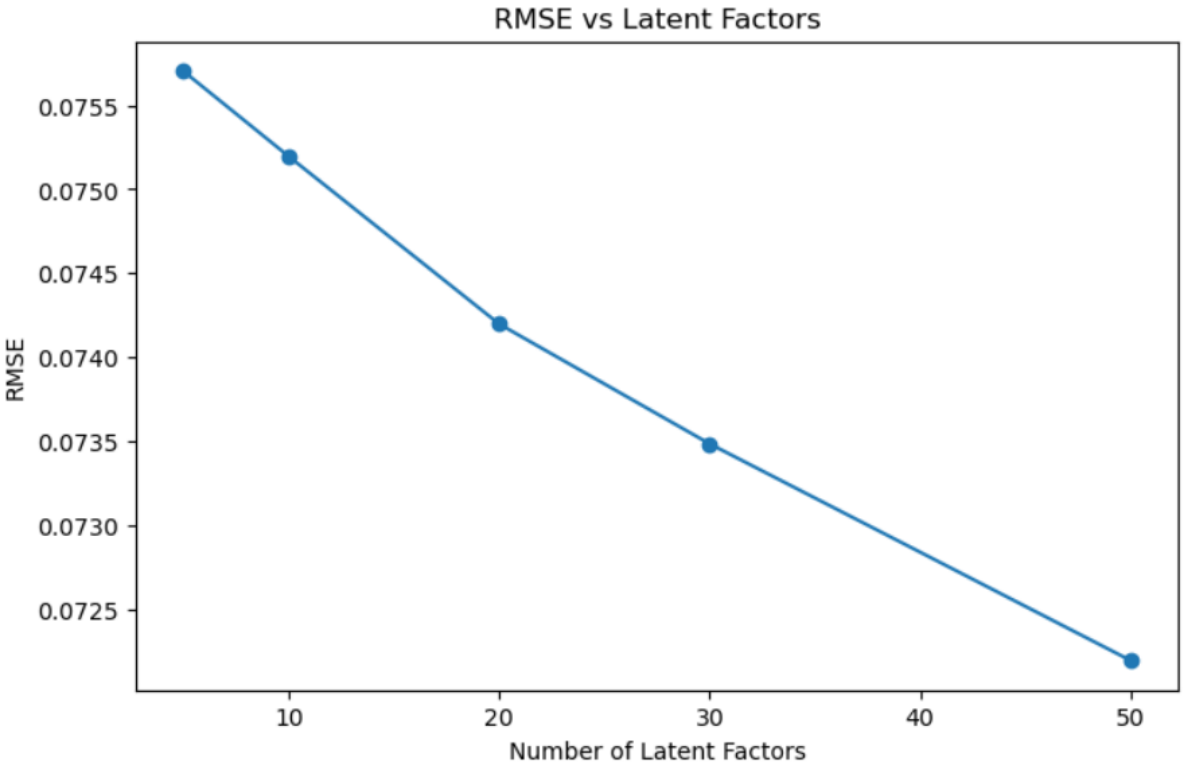
To evaluate the netflix recommendation system performance, quantitative and qualitative evaluation techniques will be combined. Error based measures like Root Mean Square Error (RMSE) will be used to determine how well the ratings is predicted.

```
# 📌 Step 11: Evaluate model (RMSE on train set)
mse = mean_squared_error(train_matrix_np.flatten(), reconstructed_matrix.flatten())
rmse = np.sqrt(mse)
print("Training RMSE:", rmse)
```

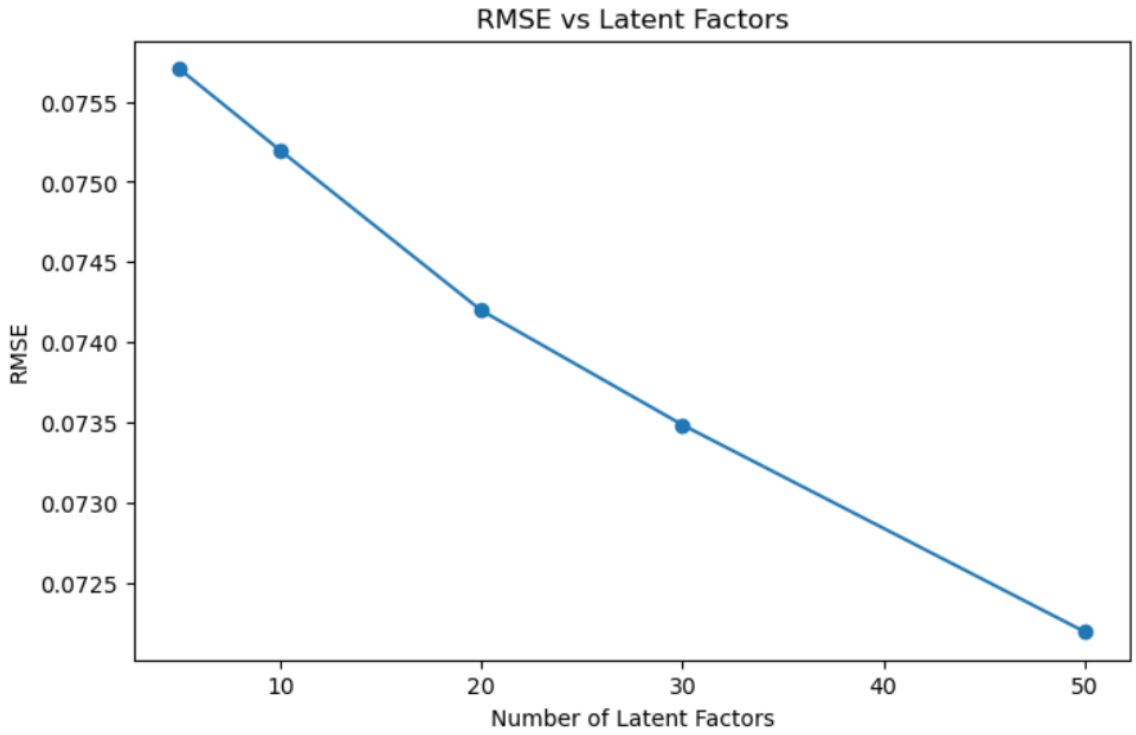
```
Training RMSE: 0.07419910075285181
```



This is the visualization of RMSE



Conclusion:



References:

1)Springer chapter in *Recommender Systems Handbook*. link.springer.com+1

https://link.springer.com/chapter/10.1007/978-0-387-85820-3_3 link.springer.com

2)Open access article: Advances in Artificial Intelligence, 2009. onlinelibrary.wiley.com

https://onlinelibrary.wiley.com/doi/10.1155/2009/421425

4)Su, X. & Khoshgoftaar, T. M. (2009). *A Survey of Collaborative Filtering Techniques.* onlinelibrary.wiley.com

Netflix Website

5)This dataset comes from the IMDB website data is collected by using web scraping