Mid Report

Movie Recommendation System

Problem Statement: The goal of this project is to design and implement a recommendation system that utilises AI algorithms to generate personalised content lists for individuals. Leveraging factors such as user profiles, search/ browsing history, demographic traits, and similar user preferences, the system aims to provide tailored recommendations to enhance user satisfaction and engagement.

Dataset: The dataset is divided into 4 different files.

1. `Links.csv` : Identifiers that can be used to link to other sources of movie data are contained in this file. It contains 3 columns, movieId (which is consistent throughout all the files), imdbId and tmdbId through which we can access the movie lens, imdb and tmdb webpage of the particular movie using id as suffix in the url.

2. `Movies.csv` : Movie information is contained in this file and each row represents one movie. It contains 3 columns, movieId, title and genres. Title of the movie also contains year in which it was released. Genres are divided into 18 different categories and one movie can have multiple.

3. `Ratings.csv` : Each row of this file represents one rating of one movie by one user. It has 4 columns, userId, movieId, rating and timestamp. Ratings are made on 5-star scale with half-star increment. Timestamp represents seconds since midnight Coordinated Universal Time (UTC) of January 1, 1970.

4. `Tags.csv` : Each row represents one tag applied to one movie by one user. It has 4 columns, userId, movieId, tag and timestamp. Tags are user generated metadata about movies. Each tag is typically a short phrase or word. The meaning, value and purpose of a particular tag is determined by each user.

Proposed approach:

After the initial data visualisation and some minor preprocessing we will merge the data of all the 4 files in single file using the consistent userId and movieId. Also to utilise the ratings of imdb, tmdb and some other features like popularity we will extract this data from the respective webpages using ids available in links.csv file.

We will create 3 different models using different classifiers like XGBRegressor, Random forest regressor and Bagging regressor.

1. Model-1 : Simple recommender, Generalised for everybody, The idea behind this model is that movies that are more popular and critically acclaimed will have a higher probability of being liked by users and will be recommended thus it does not give personalised recommendations.

2. Model-2 : This model is based on collaborative filtering it assumes that user gives similar ratings to similar movies and use this as a fundamental for recommendation.

3. Model-3 : This model provides recommendation based on the content provided by user. For example, tags.

Early Results :

We obtained the preprocessed dataset by extracting necessary information from all sites as mentioned earlier.

	title	year	vote_count	vote_average	popularity	genres	wr
15480	Inception	2010	14075	8	29.108149	[Action, Thriller, Science Fiction, Mystery, A...	7.914659
12481	The Dark Knight	2008	12269	8	123.167259	[Drama, Action, Crime, Thriller]	7.902632
22879	Interstellar	2014	11187	8	32.213481	[Adventure, Drama, Science Fiction]	7.893653
2843	Fight Club	1999	9678	8	63.869599	[Drama]	7.877957
4863	The Lord of the Rings: The Fellowship of the Ring	2001	8892	8	32.070725	[Adventure, Fantasy, Action]	7.867793
292	Pulp Fiction	1994	8670	8	140.950236	[Thriller, Crime]	7.864608
314	The Shawshank Redemption	1994	8358	8	51.645403	[Drama, Crime]	7.859864
7000	The Lord of the Rings: The Return of the King	2003	8226	8	29.324358	[Adventure, Fantasy, Action]	7.857755
351	Forrest Gump	1994	8147	8	48.307194	[Comedy, Drama, Romance]	7.856463
5814	The Lord of the Rings: The Two Towers	2002	7641	8	29.423537	[Adventure, Fantasy, Action]	7.847591
256	Star Wars	1977	6778	8	42.149697	[Adventure, Action, Science Fiction]	7.829633
1225	Back to the Future	1985	6239	8	25.778509	[Adventure, Comedy, Science Fiction, Family]	7.816099
834	The Godfather	1972	6024	8	41.109264	[Drama, Crime]	7.810081
1154	The Empire Strikes Back	1980	5998	8	19.470959	[Adventure, Action, Science Fiction]	7.809326
46	Se7en	1995	5915	8	18.457430	[Crime, Mystery, Thriller]	7.806877

```
RangeIndex: 100226 entries, 0 to 100225
Data columns (total 29 columns):
 #   Column               Non-Null Count   Dtype
---  ------               --------------   -----
 0   Unnamed: 0           100226 non-null  int64
 1   userId               100226 non-null  int64
 2   movieId              100226 non-null  int64
 3   rating               100226 non-null  float64
 4   timestamp            100226 non-null  int64
 5   title                100226 non-null  object
 6   Avg_Rating_TMDB      100226 non-null  float64
 7   Vote_Count_TMDB      100226 non-null  float64
 8   year                 100209 non-null  object
 9   (no genres listed)   100226 non-null  int64
 10  Action               100226 non-null  int64
 11  Adventure            100226 non-null  int64
 12  Animation            100226 non-null  int64
 13  Children             100226 non-null  int64
 14  Comedy               100226 non-null  int64
 15  Crime                100226 non-null  int64
 16  Documentary          100226 non-null  int64
 17  Drama                100226 non-null  int64
 18  Fantasy              100226 non-null  int64
 19  Film-Noir            100226 non-null  int64
 20  Horror               100226 non-null  int64
 21  IMAX                 100226 non-null  int64
 22  Musical              100226 non-null  int64
 23  Mystery              100226 non-null  int64
 24  Romance              100226 non-null  int64
 25  Sci-Fi               100226 non-null  int64
 26  Thriller             100226 non-null  int64
 27  War                  100226 non-null  int64
 28  Western              100226 non-null  int64
dtypes: float64(3), int64(24), object(2)
```