Movie Recommendation System

Team Members: Derrie Susan Varghese, Sayan Biswas, Sneha Agarwal, Varun Jagadeesh

Dataset

With this project, we aim to build a movie recommendation system. We are using the MovieLens 25M dataset available here. The data is present across 6 CSV files: `genome-scores.csv`, `genome-tags.csv`, `links.csv`, `movies.csv`, `ratings.csv` and `tags.csv'. The dataset contains 25 million ratings and one million tag applications applied to 62000 movies by 162,000 users. The data was created between 1995 and 2019.

For each movie, we have the title and the genre it belongs to. Movie titles are entered manually or imported from other sources and include the year of release in parentheses. Genres are a pipe-separated list and fall into 1 or multiple of the 18 categories (Action, Adventure, etc.). This data is present in the 'movies.csv' file. For each user - movie pair we have the movie rating the user has provided along with the timestamp. Ratings are made on a 5-star scale, with half-star increments (0.5 stars - 5.0 stars). Timestamps represent seconds since midnight Coordinated Universal Time (UTC) of January 1, 1970. This information is contained in the 'ratings.csv' file. All selected users have rated at least 20 movies.

For each user - movie pair we also have the associated tags (sci-fi, fiction, tense, climate, etc.) along with the timestamp. Tags are user-generated metadata about movies. Each tag is typically a single word or short phrase. The meaning, value, and purpose of a particular tag are determined by each user. This information is present in the 'tags.csv' file. We also have the tag genome data available. The tag genome is a data structure that contains tag relevance scores for movies. The tag genome encodes how strongly movies exhibit particular properties represented by tags. This information is present in 'genome-scores.csv' and 'genome-tags.csv' files. Additionally, we also have 'links.csv' file which contains identifiers that can be used to link to other sources of movie data such as the IMDB website.

Exploratory Data Analysis

From this extensive dataset, we aim to answer questions like:

1. Which genres have the maximum number of movies and how does it change with time?
2. Which genres receive the highest ratings and how do these ratings change with time?
3. What is the distribution of star ratings across genres?
4. What tags best summarize a movie genre?
5. Which is the best movie for every decade?
6. Which is the best year for a genre?
7. Genre preference for specific users?
8. What are the association rules for users watching movies?

Proposed Methodology

We would start with tidying the data (if necessary) using R such as

- Dealing with NAs
- Separating genres which have values like "Action | Comedy" into separate rows
- Timestamp conversion into a readable format
- Extracting year from the movie title

We then aim to do data visualization in R using Exploratory data analysis techniques. We plan on utilizing cosine similarity to find the similarity between any two users and any two movies. We then plan on using Market basket analysis to find association rules. Our ultimate goal of the project is to recommend new movies to users, to dig out what users may like but they did not know before.