Movie Recommendation System Using MovieLens 20M Dataset

Problem Statement - Hypothesis

The goal of this project is to develop a machine learning-based recommendation system that suggests movies to users based on their viewing history, preferences, and demographic information. The system will leverage collaborative filtering techniques, specifically Singular Value Decomposition (SVD), to predict users' ratings for movies they haven't yet watched. The hypothesis is that by analyzing past user interactions, the system can accurately predict movie preferences and improve the user experience by offering personalized recommendations.

Context

In the modern digital entertainment world, users have access to vast catalogs of movies on streaming platforms. However, discovering new content can be overwhelming. A recommendation system aims to solve this issue by providing personalized suggestions based on user preferences. By analyzing user interactions (e.g., ratings) and demographic information, the system can recommend movies that users are likely to enjoy, increasing user engagement and satisfaction.

The MovieLens 20M dataset contains 20 million user ratings for movies and includes ratings, metadata about movies (genres, release year) and basic user demographic information (age, gender, occupation).

Criteria for Success

Success will be determined based on the system’s ability to accurately predict movie ratings and provide relevant recommendations. The system will be evaluated using:

- **RMSE (Root Mean Squared Error)**: To measure the difference between actual and predicted ratings.
- **User Satisfaction**: Based on offline testing and simulations, using metrics like Precision, Recall, and Mean Reciprocal Rank (MRR).
- **Scalability**: The system should be capable of efficiently handling the large-scale MovieLens dataset without significant performance degradation.

Scope of Solution Space

The primary solution will focus on collaborative filtering techniques:

- **SVD (Singular Value Decomposition)** will be used for matrix factorization to discover latent features in the user-movie rating matrix.

- Content-based filtering techniques may be explored to analyze movie metadata, but the main focus is on collaborative filtering.
- A hybrid approach that combines collaborative filtering and content-based methods may be considered to improve prediction accuracy.

The project will also explore ways to integrate demographic information (age, gender) into the recommendation engine to improve predictions.

Constraints within Solution Space

- **Data Limitations**: While the MovieLens dataset contains rich user rating data, it lacks real-time viewing patterns (e.g., pause, stop), which limits the depth of understanding user preferences.
- **Computational Resources**: SVD can be computationally expensive for large datasets. Efficient implementations and model optimization are necessary to handle the large dataset.
- **Cold Start Problem**: For users or movies with limited data, it may be challenging to provide accurate recommendations.

Stakeholders to Provide Key Insight

- **End Users (Movie Viewers)**: Understanding their needs, preferences, and behaviors will be essential to developing a system that provides relevant recommendations.
- **Streaming Platform Subscribers**: Their feedback on integrating the system into existing infrastructure will be key, especially regarding real-time recommendation performance.

Key Data Sources

- **MovieLens 20M Dataset**: This will be the primary dataset used, containing 20 million user ratings of movies, along with demographic information and movie metadata (genres, titles, release years). This dataset is well-structured and supports collaborative filtering algorithms.
- **Additional Datasets:** If necessary, external datasets with more detailed movie metadata or real-time user interaction data can be incorporated to enhance the system (ex: https://www.kaggle.com/datasets/netflix-inc/netflix-prize-data).