# **Project 1 : Movie Recommendation System**

Tejashri Bhilare

MS in Data Science, Bellevue University

DSC680 – Applied Data Science

Professor - Amirfarrokh Iranitalab

April, 2025

**Movie Recommendation System**

**Executive Summary**

This white paper presents a Movie Recommendation System using collaborative, content-based, and demographic filtering methods. The aim is to create a personalized user experience by predicting whether a user will enjoy a particular movie based on previous interactions.

The system leverages the MovieLens dataset and addresses the common challenges associated with recommendation systems such as cold-start problems and duplicate movie listings.

**Introduction**

Recommendation systems are information filtering tools designed to present users with the most relevant content. Tech giants like Netflix, Amazon, and YouTube rely on them to increase engagement and retention.

With increasing data volumes, it's essential to tailor content delivery efficiently. This paper explores the methodologies for building a robust recommendation system using the MovieLens dataset.

**Problem Statement**

Video content platforms face the challenge of content overload. Users struggle to find content that matches their tastes. The business need is a system that understands user preferences and suggests relevant movies to enhance user satisfaction and platform retention.

**Proposed Solution**

We propose a hybrid recommendation system that integrates:

1. Demographic Filtering – Recommends popular content among similar demographic groups.
2. Content-Based Filtering – Uses item metadata (genre, director, actors) to recommend similar movies.
3. Collaborative Filtering – Finds users with similar tastes and recommends what they liked.

These methods combined provide a personalized and scalable recommendation engine.

**Dataset**

**MovieLens Dataset:** [**https://grouplens.org/datasets/movielens/latest/**](https://grouplens.org/datasets/movielens/latest/)

* 100,000+ ratings
* 3,000+ tag applications
* 9,000+ movies
* 600+ users (1995-2018)
* Anonymized user data

**System Design & Approach**

1. Data Preprocessing:

* Cleaning nulls and duplicates
* Merging ratings with movie metadata
* Encoding genres

2. Content-Based Filtering:

* TF-IDF vectorization of movie descriptions
* Cosine similarity to find similar movies

3. Collaborative Filtering:

* Matrix factorization using Singular Value Decomposition (SVD)
* User-Item interaction matrix

4. Hybrid Model:

* Weighted ensemble of collaborative and content-based scores.

**Evaluation Metrics**

To evaluate model performance, the following metrics will be used:

* **RMSE (Root Mean Square Error)**: Measures error between actual and predicted ratings
* **Precision@k**: Measures the proportion of recommended items that are relevant
* **Recall@k**: Measures the proportion of relevant items recommended

**Ethical Implications**

* **User Privacy**: MovieLens dataset is anonymized. No personal or demographic identifiers.
* **Bias in Recommendations**: Popularity bias and overfitting are addressed through algorithm diversification.
* **Transparency**: The methodology and evaluation are documented for reproducibility.

**Challenges**

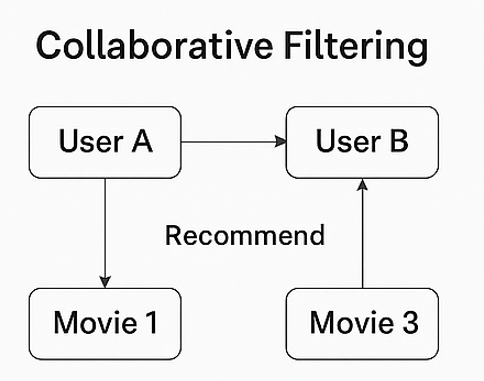
* **Cold Start Problem**: New users/movies have no prior data.
* **Duplicate Titles**: Movies with multiple titles can affect accuracy.
* **Scalability**: Processing time increases with user-item matrix size.

**Conclusion**

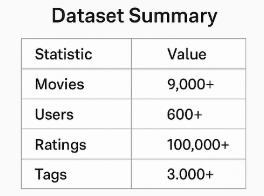
A movie recommendation system based on hybrid filtering methods can effectively enhance user experience on content platforms. Using the MovieLens dataset, this system leverages both metadata and user behavior to provide accurate suggestions. Ethical considerations and scalability remain central to its development.

**Illustrations**

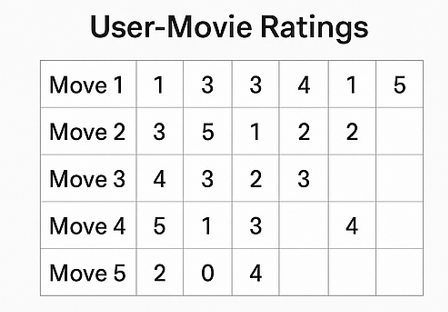
1. Architecture diagram of hybrid recommendation system



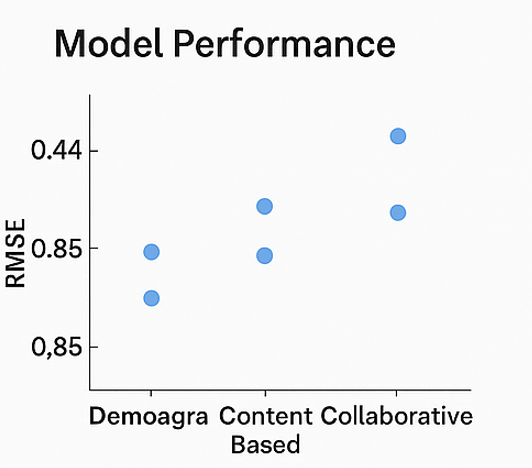
1. Sample ratings matrix with user and movie IDs



1. Genre encoding example

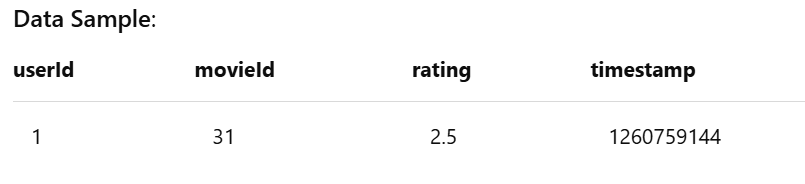


1. RMSE score comparison of algorithms



**Appendix**

* **Data Sample**:



* **Tools/Libraries**: Python, Pandas, Scikit-learn, Surprise, Matplotlib
* **Data Dictionary**:
  + userId: anonymized ID
  + movieId: unique movie identifier
  + rating: user-assigned score (0.5-5.0)
  + timestamp: Unix time of rating

**References**

1. Great Learning. (n.d.). *Masterclass on Movie Recommendation System*. Retrieved April 2025, from https://www.mygreatlearning.com/blog/masterclass-on-movie-recommendation-system/

2. Koren, Y. (2008). Factorization meets the neighborhood: A multifaceted collaborative filtering model. *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 426–434. https://doi.org/10.1145/1401890.1401944

3. GroupLens. (n.d.). *MovieLens latest datasets*. Retrieved April 2025, from <https://grouplens.org/datasets/movielens/latest/>

4. NVIDIA. (n.d.). *Recommendation system*. Retrieved April 2025, from <https://www.nvidia.com/en-us/glossary/recommendation-system/>