MOVIELENS RECOMMENDER SYSTEM

Our goal is to build a system that recommends the top 5 movies to a specific user based on their historical ratings.

The core idea is to leverage the collective wisdom of all users in the MovieLens dataset to predict what an individual will enjoy.

This problem is tackled by answering four fundaments questions:

01

02

03

1. How do we find users with the same movie taste? (Collaborative Filtering (User-User)

2.How do we find movies that are similar to the ones the user already likes? (Content-Based/item-item Collaborative Filtering)

3. How can we predict a user's rating for an unrated movie?

4. How do we generate the final Top 5 recommendations?

Movies Dataset Overview

Movies Data

- Entries: 9,742 movies

Ratings Data

- Entries: 100,836 ratings

ADDITIONAL DATASET

Links dataset

Ratings dataset

Baseline Model

A simple collaborative filtering baseline that predicts a rating by combining three core components:

1. **Global Mean (μ): The average rating across the entire dataset**
2. **User Bias (b_u): How much a specific user tends to rate above/below the global mean**
3. **Movie Bias (b_i): How much a specific movie is rated above/below the global mean**

Sample Recommendations (User 429)

User's Taste Profile:

Rated war/action (Crimson Tide), space adventure (Apollo 13), and animation (Aladdin) all as 5 stars.

Top 5 Recommendations:

All predicted as perfect 5.0-star matches:

1. **Paths of Glory** (1957) – War/Drama
2. **Jules and Jim** (1961) – Drama/Romance
3. **Yojimbo** (1961) – Action/Drama
4. **Mr. Death** (1999) – Documentary
5. **Five Easy Pieces** (1970) – Drama

SVD Model (Singular Value Decomposition)

Overview

A collaborative filtering technique that decomposes the sparse user-item rating matrix into lower-dimensional latent factor matrices to uncover hidden patterns and preferences.

Why It's Used

Captures Latent Features: Automatically discovers hidden dimensions (e.g., "genre preference", "movie style") without explicit labels

Handles Sparsity: Works well with sparse rating matrices by learning from patterns across users/items

Enables Personalization: Predicts ratings for unseen user-item pairs via dot products of latent vectors

SVD Recommendations

The Singular Value Decomposition (SVD) model predicts ratings based on latent factors from user-item interactions.

Schindler’s List (1993) – High predicted rating (4.45) aligns with user's preference for drama and historical themes.

Shawshank Redemption (1994) – Classic drama with strong narrative, similar to user’s top-rated films.

Blade Runner 2049 (2017) – Sci-Fi pick, possibly influenced by user’s interest in Apollo 13 (space/tech themes).

Failure to Launch (2006) – Comedy/Romance outlier; may reflect latent genre diversity in user’s history.

Casablanca (1942) – Timeless drama/romance; model likely captures classic film appeal.

HYBRID MODEL

We' combined three approaches:

1. SVD Collaborative Filtering (for users with sufficient ratings)
2. Content-Based Filtering(for cold start using movie genres)

3.Baseline Model(fallback for cases where both fail) This approach addresses the cold start problem while maintaining the accuracy of collaborative filtering for established users

Hybrid Recommendation Testing

The hybrid model was tested on two distinct user profiles to demonstrate its adaptive behavior:

1. Cold-Start User (User ID: 710)

Characteristics: New user with no rating history → classic cold-start problem.

Hybrid Recommendations:

All recommended movies have the same score (3.537), indicating:

Content-based or baseline components are dominant

Recommendations are popular, high-rated films (Forrest Gump, Shawshank Redemption, Pulp Fiction, etc.)

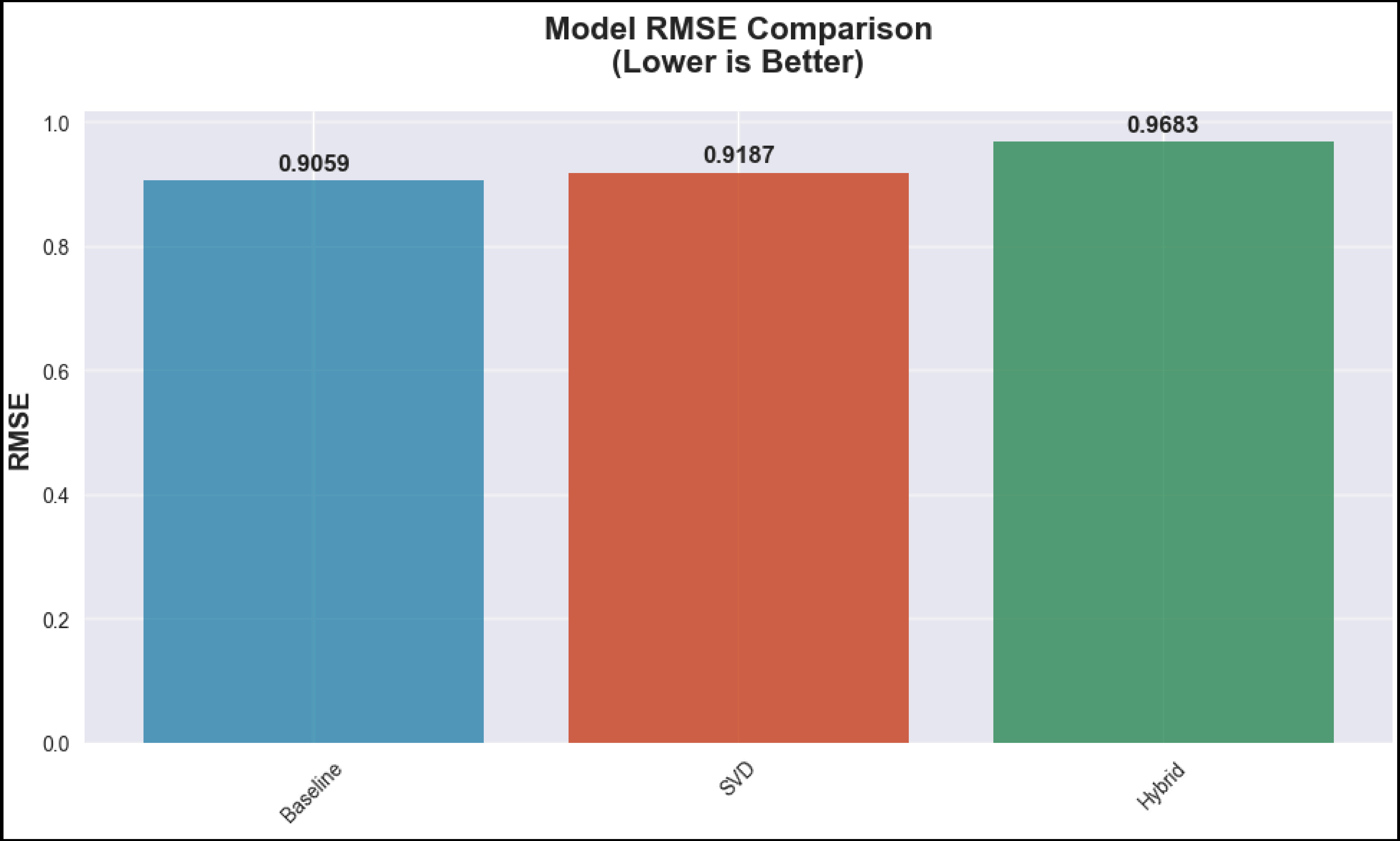
Genres are diverse, providing a broad starting point for user preference discovery

Strategy Applied:

Uses fallback mechanisms (content-based + baseline)

Avoids empty recommendations

Introduces user to widely-acclaimed titles


Model RMSE Comparison
(Lower is Better)
RMSE
1.0
0.8
0.6
0.4
0.2
0.0
0.9059
0.9187
0.9683
Baseline
SVD
Hybrid

Explanation

Root Mean Square Error (RMSE) measures prediction accuracy, where lower values are better. The Baseline model has the lowest RMSE (0.9059), meaning its predictions are closest to the actual values on average. Both SVD (0.9187) and the Hybrid model (0.9683) performed worse, with the Hybrid model having the highest error.

Conclusion

The Baseline model outperformed the more complex SVD and Hybrid models in this evaluation. Adding complexity (via matrix factorization in SVD or combining methods in the Hybrid approach) did not improve predictive accuracy under the RMSE metric.

Recommendation

1. Use the baseline model for deployment, as it provides the best accuracy.
2. Investigate why SVD and Hybrid models underperformed—potential issues could include:
3. Insufficient tuning of hyperparameters.
4. Data characteristics that do not suit factorization or hybrid approaches well.
5. Overfitting in more complex models that RMSE captured in test results.
6. Consider validating with additional metrics (e.g., MAE, precision@k) to ensure the Baseline also meets other business or recommendation-quality goals.

Thank You