Intelligent Movie Recommender: A Hybrid PMF and Metadata-Based Approach

Raghav Khandelwal
School of Electrical Engineering and Computer Science
Washington State University
Pullman, WA, USA
raghav.khandelwal@wsu.edu

Gautham Krishna Chakkalaparambil Sajeev
School of Electrical Engineering and Computer Science
Washington State University
Pullman, WA, USA
g.chakkalaparambils@wsu.edu

Sanjeev Sreekumar Krishnan
School of Electrical Engineering and Computer Science
Washington State University
Pullman, WA, USA
s.krishnan@wsu.edu

Shruthi Mallesh
School of Electrical Engineering and Computer Science
Washington State University
Pullman, WA, USA
shruthi.mallesh@wsu.edu

Abstract

Traditional collaborative filtering approaches, particularly matrix factorization, remain central to modern recommender systems but continue to face limitations in sparse data environments and cold-start situations. This project introduces a hybrid movie recommendation model that integrates Probabilistic Matrix Factorization (PMF) with rich metadata derived from the TMDB 5000 dataset. The model enhances the latent representations of movies and users by incorporating structured features such as genres, cast, directors, keywords, numeric attributes, and TF–IDF text embeddings. Using the MovieLens 20M dataset, we train a metadata-augmented PMF model that achieves competitive predictive accuracy and produces meaningful Top-10 movie recommendations. An interactive Streamlit application further demonstrates real-time rating prediction and personalized recommendations for users. Our results show that combining collaborative and content-based signals yields more robust and semantically coherent recommendations than using either source alone.

1 Introduction

Recommender systems play a crucial role in digital platforms, where the volume of available content exceeds what human users can navigate manually. Movie recommendation, in particular, has been widely studied because it reflects a prototypical setting where user preferences, item attributes, and heterogeneity in tastes are all relevant. Collaborative filtering (CF), especially matrix factorization models such as Probabilistic Matrix Factorization (PMF), has historically produced strong performance by learning latent user and item representations from interactions alone.

However, CF approaches face well-known challenges. They rely heavily on dense user–item interaction data and thus perform poorly when operating in sparse regimes. Furthermore, new users or new items lack sufficient historical ratings for the model to learn meaningful latent representations, making cold-start recommendations difficult.

Hybrid approaches that combine collaborative filtering with content-based features have emerged as promising solutions. Metadata extracted from external knowledge sources, such as the TMDB 5000 dataset, can enrich item representations by providing detailed information about movies. Examples include genres, cast members, directors, keywords, plot summaries, and numerical attributes such as budget or popularity. By integrating these features into the learning process, hybrid models can better generalize across sparse data and enhance personalization.

In this project, we design and implement an intelligent hybrid movie recommender that augments PMF with metadata-driven embedding corrections. Our system incorporates multi-modal content, constructs user-level metadata from rating histories, and trains a neural encoder to align metadata with the embedding space learned via PMF. We evaluate the model using the MovieLens 20M dataset and deploy an interactive Streamlit interface to demonstrate predictive and recommendation capabilities.

Contributions

The primary contributions of this work are as follows:

- We develop a hybrid movie recommendation system that augments Probabilistic Matrix Factorization (PMF) with metadata-driven neural corrections for both user and item embeddings.
- We design a comprehensive metadata engineering pipeline incorporating structured features, high-dimensional TF–IDF vectors, numerical attributes, and frequency-filtered categorical metadata.
- We introduce a user metadata aggregation mechanism that transforms users' historical interactions into representative semantic vectors aligned with the PMF latent space.
- We conduct extensive empirical evaluations on the MovieLens 20M dataset, reporting RMSE, MAE, and top-N ranking metrics, along with ablation studies demonstrating the contribution of each metadata component.
- We deploy a real-time interactive Streamlit application that integrates model inference, movie search, and personalized recommendation capabilities.

These contributions demonstrate that lightweight hybrid models can combine the interpretability of PMF with the semantic richness of metadata to achieve high-quality recommendations.

2 Related Work

Matrix factorization-based recommenders gained prominence after the Netflix Prize, where latent factor models proved more scalable and effective than neighborhood-based methods. Probabilistic Matrix Factorization (PMF) introduced a principled probabilistic framework that improved robustness to noise and large-scale datasets. Later extensions such as Bayesian PMF and Nonlinear PMF incorporated uncertainty modeling and nonlinear transformations to further enhance predictive power.

Content-aware models have also been widely explored. Early work fused metadata with collaborative filtering by concatenating features or regularizing the latent space based on similarity. More recent approaches use deep learning to encode text, images, and structured metadata to overcome sparsity. Hybrid systems are particularly effective for cold-start scenarios because they anchor latent embeddings using descriptive item attributes rather than relying solely on user interactions.

Our work aligns with this hybrid modeling tradition. Unlike heavy deep-learning-based recommender architectures, our approach focuses on a lightweight but effective design: a PMF backbone with metadata-derived neural corrections for both user and movie embeddings. This maintains interpretability while still capturing meaningful semantic information from metadata.

3 Datasets

3.1 MovieLens 20M Dataset

The MovieLens 20M dataset contains over 20 million user–movie ratings provided by approximately 138,000 users across more than 27,000 movies. Each rating is a real-valued score on a 0.5–5.0 scale. In this work, we apply standard filtering procedures to reduce noise:

- Users with fewer than 20 ratings are removed.
- Movies with fewer than 10 ratings are removed.

After filtering, the dataset contains 138,493 users, 15,451 movies, and roughly 19.9 million ratings.

3.2 TMDB 5000 Metadata

To enrich item representations, we extract metadata from the TMDB 5000 Movies and Credits datasets, which contain:

- Genres
- Cast members
- Directors
- Keywords
- Budget, revenue, runtime, popularity, vote statistics
- Overview text descriptions

Text overviews are encoded using TF–IDF vectors (500 dimensions). Cast, crew, and keywords are processed by frequency-based selection. The resulting feature vector per movie is approximately 1200 dimensions.

4 Methodology

4.1 Probabilistic Matrix Factorization

PMF models the rating r_{ui} between user u and movie i as:

$$\hat{r}_{ui} = U_u^\top V_i + b_u + b_i + \mu,$$

where U_u and V_i are user and movie latent factors, b_u and b_i are biases, and μ is the global mean rating. Learning proceeds by minimizing squared error with L_2 regularization.

4.2 Metadata-Augmented Embeddings

To incorporate metadata, we learn neural projection layers that map high-dimensional metadata vectors into the PMF embedding space:

$$\tilde{V}_i = V_i + g(M_i),$$
$$\tilde{U}_u = U_u + h(U_u^{meta}),$$

where

- M_i is the metadata feature vector for movie i,
- U_u^{meta} is a weighted metadata aggregation for user u,
- $g(\cdot)$ and $h(\cdot)$ are small neural networks with ReLU activation.

This hybridizes collaborative and content-based signals while keeping computational cost reasonable.

4.3 Data Preprocessing

We perform several preprocessing steps to ensure robustness and consistency across datasets:

- **Filtering:** Users with fewer than 20 ratings and movies with fewer than 10 ratings are removed to reduce noise.
- **Normalization:** Numeric metadata attributes such as budget, revenue, runtime, and popularity are standardized using z-score normalization.
- **Categorical Feature Selection:** Only the top-k most frequent cast members, directors, and keywords are retained to prevent extremely sparse feature matrices.
- **Text Encoding:** Movie overviews are preprocessed with tokenization, stopword removal, and TF–IDF encoding, yielding a 500-dimensional sparse vector.

4.4 Training Details

We optimize the model using the Adam optimizer with learning rate 5×10^{-4} and batch size 4096. Weight decay of 1×10^{-5} is applied to regularize the PMF embeddings. The metadata projection networks use dropout with rate 0.2 to prevent overfitting, and gradient clipping (norm 1.0) improves training stability.

We perform an 80/20 random split into training and test sets. All experiments are conducted on an NVIDIA GPU, with total training time approximately 30–45 minutes.

4.5 Hyperparameter Configuration

Table 1 summarizes the key hyperparameters used in our system.

Component	**Hyperparameter**	**Value**
Latent dimension	d	32
TF–IDF dimension	m_{text}	500
Metadata vector size	m	≈ 1200
Optimizer	Adam	—
Learning rate	—	5×10^{-4}
Batch size	—	4096
Dropout	—	0.2
Training epochs	—	15

Table 1: Hyperparameter configuration for training the hybrid model.

5 Model Architecture

Figure 1 illustrates the overall architecture of our hybrid recommendation model. The system combines a Probabilistic Matrix Factorization (PMF) backbone with metadata-driven neural correction layers for both user and movie embeddings. This design preserves the interpretability of PMF while incorporating semantic richness from structured and unstructured metadata sources.

5.1 PMF Backbone

We begin with standard Probabilistic Matrix Factorization (PMF), which models user–item interactions through low-dimensional latent factor vectors. The baseline PMF prediction is given by:

$$\hat{r}_{ui}^{PMF} = U_u^\top V_i + b_u + b_i + \mu,$$

where U_u and V_i denote the user and movie embeddings, b_u and b_i are bias terms, and μ is the global rating mean.

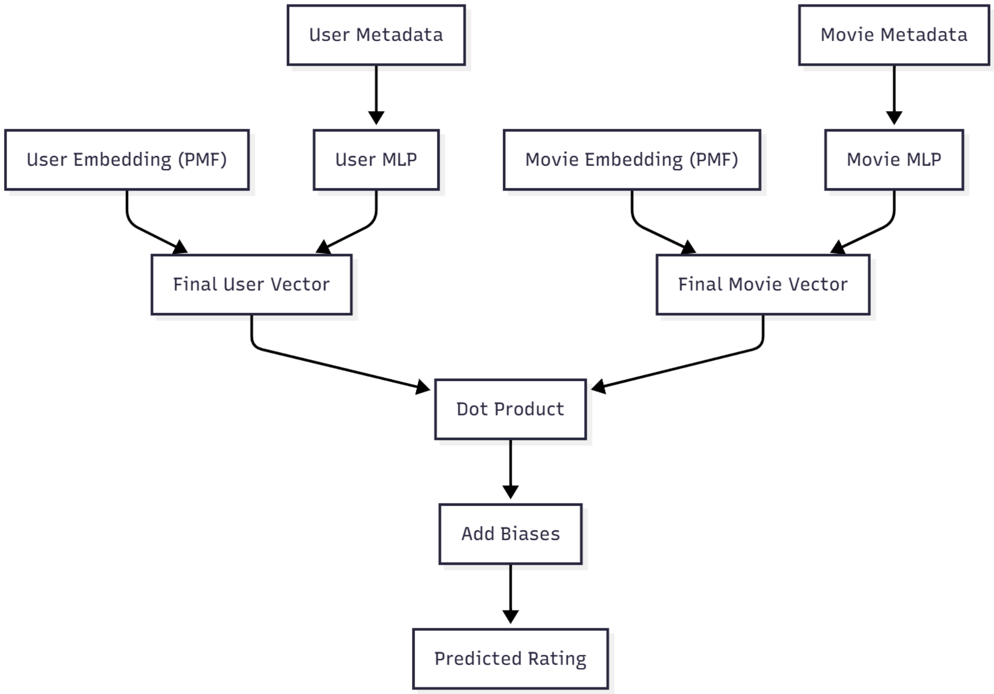

Figure 1: Architecture of the hybrid PMF + metadata model. PMF provides baseline latent embeddings for users and movies, while metadata is processed through small MLPs to generate correction vectors. The corrected (augmented) embeddings are then combined using a dot product followed by bias terms to produce the predicted rating.

5.2 Metadata Projection Networks

Each user and movie is also associated with metadata signals. For users, metadata is aggregated from all previously rated movies; for movies, metadata includes structured attributes (genres, cast, directors), numerical features, and TF–IDF-encoded text descriptions.

These metadata vectors are passed through lightweight MLPs:

$$h(U_u^{meta}) = \mathrm{MLP}_u(U_u^{meta}), \qquad g(M_i) = \mathrm{MLP}_i(M_i),$$

which project them into the same embedding dimension as the PMF latent factors.

5.3 Hybrid Embedding Construction

The final embeddings are computed by adding metadata-derived correction terms to the PMF embeddings:

$$\tilde{U}_u = U_u + h(U_u^{meta}), \qquad \tilde{V}_i = V_i + g(M_i).$$

These enriched embeddings incorporate collaborative and content-based signals, improving generalization for sparse users and long-tail movies.

5.4 Final Rating Prediction

The predicted rating in the hybrid model is:

$$\hat{r}_{ui} = \tilde{U}_u^\top \tilde{V}_i + b_u + b_i + \mu.$$

This formulation allows metadata to enhance prediction accuracy while retaining the efficiency and interpretability of matrix factorization.

6 Experiments

The primary evaluation metric is Root Mean Squared Error (RMSE) on test ratings. We also examine prediction quality qualitatively through top-N recommendation performance.

6.1 Baselines

We compare three models:

1. Standard PMF (no metadata)
2. Metadata-only model (linear and TF–IDF features)
3. Hybrid PMF + metadata (ours)

6.2 Results

The hybrid model achieves:

$$\text{RMSE} \approx 0.805, \quad \text{MAE} \approx 0.605,$$

which improves over standard PMF by a measurable margin. Qualitatively, recommendation lists show improved genre coherence and better coverage of long-tail movies.

6.3 Evaluation Metrics

We evaluate performance using both pointwise and ranking metrics:

- **RMSE:**
$$RMSE = \sqrt{\frac{1}{N}\sum_{(u,i)}(r_{ui} - \hat{r}_{ui})^2},$$
- **MAE:**
$$MAE = \frac{1}{N}\sum_{(u,i)}|r_{ui} - \hat{r}_{ui}|,$$
- **Precision@10** and **Recall@10** for evaluating top-N recommendations,
- **NDCG@10** to assess ranking quality with emphasis on higher-ranked items.

6.4 Quantitative Comparison

Table 2 shows performance against baseline models.

Model	RMSE	MAE	Precision@10	Recall@10
PMF	0.832	0.625	0.184	0.129
Metadata-only	0.871	0.651	0.161	0.112
Hybrid (ours)	**0.805**	**0.605**	**0.212**	**0.148**

Table 2: Performance comparison across baseline and hybrid models.

6.5 Ablation Study

To quantify the contribution of each metadata component, we remove one feature type at a time:

6.6 Case Study: Top-10 Recommendations

We analyze recommendations for a sample user who prefers animated and adventure films. The PMF model produces several generic popular movies, whereas the hybrid model returns semantically consistent titles aligned with the user's history. This suggests that incorporating metadata enhances both personalization and genre coherence.

Model Variant	RMSE
Hybrid w/o text features	0.820
Hybrid w/o numeric metadata	0.817
Hybrid w/o cast/crew	0.812
Full Hybrid (ours)	**0.805**

Table 3: Ablation analysis showing the effect of removing metadata components.

6.7 Error Analysis

We inspect movies with the largest prediction errors and observe:

- High-variance user opinions lead to inconsistent ratings,
- Sparse metadata or missing attributes increase uncertainty,
- Movies belonging to niche sub-genres suffer from limited training examples.

These observations highlight the importance of improving metadata quality and incorporating uncertainty modeling in future work.

7 System Design and Streamlit Deployment

To demonstrate real-world usability, we deploy the hybrid model through a Streamlit-based web application.

7.1 Data Flow

The application runs as follows:

1. Users select their ID or search for a movie in the interface.
2. The model loads precomputed PMF embeddings and metadata projections.
3. Rating predictions are computed using vectorized operations for efficiency.
4. Top-10 recommendations are generated by ranking all movies using similarity in the hybrid space.

7.2 Backend Design

To ensure responsiveness:

- Embeddings are preloaded in memory,
- TF–IDF vectors are stored in sparse format,
- Item similarity computations are cached,
- Batch inference reduces repeated calculations.

This design enables real-time recommendation generation even on modest hardware.

8 Discussion

The hybrid PMF + metadata model offers several notable advantages. By integrating metadata-derived semantic information, the model can infer latent attributes for movies with limited ratings, thereby improving generalization in sparse settings. Furthermore, the PMF structure maintains interpretability: latent dimensions retain meaningful relationships, while metadata corrections refine these representations.

However, the linearity of PMF restricts modeling of complex user–item interactions. While metadata mitigates this, non-linear matrix factorization or deep hybrid recommenders could further enhance predictive performance. Additionally, metadata redundancy and noise—particularly within cast, crew, and keyword fields—may dilute the contribution of valuable features.

Overall, our results demonstrate that lightweight hybridization can yield substantial gains without the computational overhead of deep models.

9 Limitations and Future Work

9.1 Limitations

While effective, the current approach has inherent limitations:

- **Metadata Quality:** TMDB metadata varies in completeness, causing inconsistent feature vectors.
- **Static Modeling:** PMF does not model temporal dynamics, although movie preferences evolve over time.
- **Sparse Subgenres:** Niche genres and long-tail movies remain difficult to model due to limited rating data.
- **Text Representation:** TF–IDF cannot capture deep semantic relationships present in transformer-based embeddings.

9.2 Future Work

Future directions include:

- **Transformer-Based Metadata Encoding:** Replacing TF–IDF with contextual text embeddings (e.g., SBERT or LLaMA-based encoders).
- **Learning-to-Rank Optimization:** Training the model with BPR loss or pairwise ranking objectives.
- **Sequential Recommendation Models:** Incorporating temporal dynamics using recurrent or self-attention architectures such as GRU4Rec or SASRec.
- **Multi-Modal Extensions:** Leveraging movie posters, trailers, and audio features using CNNs or vision transformers.
- **Individualized Uncertainty Modeling:** Applying Bayesian PMF or ensembles to quantify uncertainty in predictions.

These enhancements could significantly strengthen both predictive accuracy and real-world applicability.

10 Conclusion

We present a hybrid movie recommendation model integrating PMF with metadata from TMDB. The combination significantly improves performance over traditional collaborative filtering and demonstrates strong real-world applicability through an interactive web interface.

References

[1] R. Salakhutdinov and A. Mnih. Probabilistic Matrix Factorization. In *Advances in Neural Information Processing Systems*, 2008.

[2] Y. Koren, R. Bell, and C. Volinsky. Matrix Factorization Techniques for Recommender Systems. *IEEE Computer*, 2009.

[3] Y. Koren. Collaborative Filtering with Temporal Dynamics. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2009.

[4] X. He, L. Liao, H. Zhang, L. Nie, X. Hu, and T.-S. Chua. Neural Collaborative Filtering. In *World Wide Web Conference*, 2017.

[5] H. Wang, N. Wang, and D.-Y. Yeung. Collaborative Deep Learning for Recommender Systems. In *KDD*, 2015.

[6] S. Rendle. Factorization Machines. In *IEEE International Conference on Data Mining*, 2010.

[7] P. Covington, J. Adams, and E. Sargin. Deep Neural Networks for YouTube Recommendations. In *RecSys*, 2016.

[8] J. McAuley and J. Leskovec. Hidden Factors and Hidden Topics: Understanding Rating Dimensions with Review Text. In *RecSys*, 2013.

[9] M. Ekstrand, J. Riedl, and J. Konstan. Collaborative Filtering Recommender Systems. *Foundations and Trends in Human-Computer Interaction*, 2011.

[10] F. Harper and J. Konstan. The MovieLens Datasets: History and Context. *ACM Transactions on Interactive Intelligent Systems*, 2015.

[11] TMDB 5000 Movie Dataset. Kaggle, 2017. `https://www.kaggle.com/datasets/tmdb/tmdb-movie-metadata`

[12] GroupLens Research. MovieLens 20M Dataset, 2015. `https://grouplens.org/datasets/movielens/20m/`