Movies Recommendation and Rating Prediction

Martim Teixeira
mateixeira@ucsd.edu
University of California – San Diego, Rady School of Management

Abstract

The objective of this project, part of my Web Mining and Recommender Systems course, is to design a movie recommendation system. Leveraging the principles and methodologies explored in class, my aim is to develop a system that accurately suggests movies to users based on their preferences and viewing history.

I chose to focus on movie recommendations, driven by a personal interest in this area. This decision led me to seek out a dataset that was rich in user ratings and movie preferences, a key first step in the project. Selecting the right dataset was essential, as it would lay the groundwork for all subsequent analysis. My aim in this exploratory phase is to delve into the nuances of user behavior and preferences in film, examining elements like rating trends, genre popularity, and engagement patterns. These insights will be crucial in crafting an effective movie recommendation system, reflecting my chosen area of focus.

The heart of the project is the construction of the recommender system itself. This involves applying machine learning techniques to predict user preferences and suggest movies accordingly. Additionally, the system is designed to function as a rating predictor, estimating the ratings users would likely give to movies based on their past preferences and ratings. This predictive capability enhances the recommendation process, allowing for more personalized and accurate suggestions. The model's success will be gauged by its accuracy in reflecting users' tastes and preferences.

Throughout the report, I will detail my approach to data processing, feature extraction, and model development. I will also benchmark the performance of my model against standard metrics to ensure its effectiveness. Additionally, I plan to review relevant literature, comparing my approach and findings with existing models and techniques in the field.

In the following sections, I will present a comprehensive narrative of my process in creating a movie recommendation system, highlighting both the challenges encountered and the insights gained.

Dataset

Structure

For this analysis I used the MovieLens dataset, sourced from GroupLens Research, specifically the version recommended for education and development. The MovieLens dataset is a rich collection of movie ratings, extensively used in the field of recommender systems and machine learning for its reliability and comprehensive nature.

The dataset features an extensive collection of movies spanning various genres and eras, accompanied by user ratings on a defined scale that reflects viewers' preferences and sentiments toward different movies. The data entries include a unique user ID, movie ID, the rating given, and a timestamp of when the rating occurred, paving the way for a detailed examination of user preferences and movie popularity over time.

Content and Use of Files

The dataset files are organized as comma-separated values (CSV) with a single header row to outline the content structure.

The MovieLens dataset features anonymized user IDs and movie IDs, ensuring privacy and consistency across the different data

files: ratings.csv, tags.csv, movies.csv, and links.csv. Each user and movie ID is unique, allowing for accurate cross-referencing throughout the dataset.

In the ratings.csv file, each line records a user's rating for a movie, provided on a 5-star scale with half-star increments, along with a timestamp in UTC. The data is sorted first by user ID and then by movie ID.

The tags.csv file contains user-generated tags for movies, structured in a similar format to the ratings file, including timestamps. These tags serve as descriptive metadata and are sorted in the same manner as the ratings.

Lastly, the movies.csv file details each movie's ID, title with the release year, and genres, which are listed in a pipe-separated format. Titles and genres are obtained from verified sources or entered manually, contributing to a rich categorization system within the dataset.

Interesting Findings

An initial exploration of the dataset reveals several interesting trends and patterns. For example, if we look to the distribution of movie ratings at the Figure 1, It is evident from the chart that the dataset contains a higher number of ratings towards the upper end of the scale, with 4 being the most common rating, followed by 3 and 5. This suggests a tendency among users to rate movies positively.

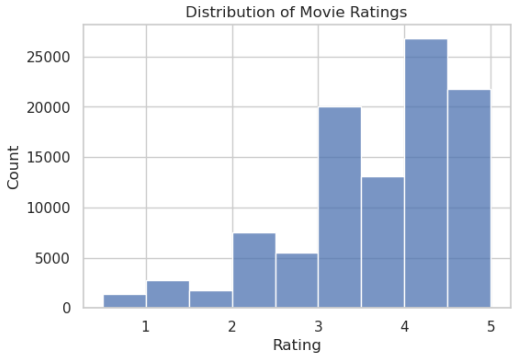

Figure 1 - Distribution of Movie Ratings

Observing now the most popular movies, using as consideration the number of ratings, the five most popular movies are:

- Forrest Gump (1994)
- Shawshank Redemption, The (1994)
- Pulp Fiction (1994)
- Silence of the Lambs, The (1991)
- Matrix, The (1999)

We are now concentrating on our users and have tallied the number of reviews each has contributed. This analysis has revealed the top five users with the highest number of reviews. They are identified by their user IDs: 414, 599, 474, 448, and 274.

To finish this initial exploratory analysis, let's consider the genre distribution within the dataset, as depicted in the Figure 2. This horizontal bar chart illustrates the number of movies across different genres. Drama leads the count significantly, followed by Comedy and Thriller, indicating a strong representation of these genres. Action and Romance also show substantial numbers, while genres such as IMAX, Western, and Film-Noir are among the least represented in the dataset.

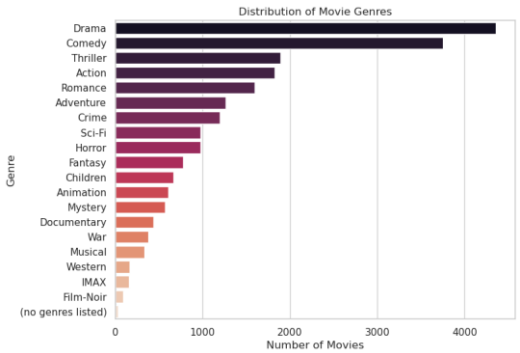

Figure 2 - Distribution of Movie Genres

The prominence of certain genres over others can reflect broader industry trends or a bias in the dataset's compilation. Notably, the category for 'no genres listed' suggests the presence of films that do not fit into traditional genre classifications or may lack genre information. This distribution is fundamental to understanding the diversity of the dataset and can have implications for the development of genre-specific recommendation algorithms.

Predictive Task

Objective

The aim of this project is to develop a recommendation system that suggests a list of 10 movies a user is likely to enjoy. This objective is realized by predicting the ratings a user might assign to various movies, which are then used to make recommendations. The methodology combines user-item interaction data with advanced collaborative filtering techniques.

Methodology

Initial Approach

The first stage in the recommendation process involves using a content-based filtering method with cosine similarity scores. This approach calculates the resemblance between movies based on their content profiles, identifying films similar to those a user has previously rated highly. A correlation matrix quantifies these similarities, and the top movies with the highest cosine similarity scores are initially considered for recommendations.

Advanced Rating Prediction

After the initial stage, the project employs two sophisticated collaborative filtering algorithms to enhance rating predictions:

- Non-negative Matrix Factorization (NMF): This method factorizes the user-item rating matrix into non-negative components, uncovering latent relationships crucial for predicting ratings.
- Singular Value Decomposition (SVD): SVD decomposes the user-item rating matrix into singular vectors, capturing underlying factors that influence user ratings.

Movies Recommendation

The recommendation process in this project begins with using Cosine Similarities to identify movies similar to those a user has previously enjoyed, focusing on content similarity for initial recommendations.

It then progresses to Non-negative Matrix Factorization (NMF) and Singular Value Decomposition (SVD) for more complex rating predictions, analyzing user-item interactions.

Finally, the model with the highest accuracy, whether Cosine Similarities, NMF, or SVD, is selected to recommend the top 10 movies to the user.

Model Evaluation and Validation

To assess the accuracy of Cosine Similarities, Non-negative Matrix Factorization (NMF), and Singular Value Decomposition (SVD), the mean squared error (MSE) was the measure selected. This measures the average squared differences between actual user ratings and the models' predictions, with lower scores indicating better accuracy.

The dataset is divided into training and testing sets to validate the models against new, unseen data, ensuring their robustness and generalizability.

The most effective model is then selected for personalized and diverse movie recommendations, enhancing the overall relevance and quality.

Model Description

As referred to in the previous section, three distinct models were implemented in this project to predict user movie ratings and make recommendations, using unique approach to analyze user-item interactions and predict user preferences, offering insights into the most effective methods for movie recommendation.

Cosine Similarities

The first model applies Cosine Similarities to a user-item matrix created from the ratings_movies_df dataset. This matrix is transformed to fill missing values with zeros, enabling the calculation of similarities between users based on their rating patterns. The similarity matrix is then used to predict ratings and recommend movies closely aligned with a user's past preferences. The

model is evaluated using mean squared error (MSE) to gauge its predictive accuracy. In this approach, the MSE was found to be 12.24.

Additionally, a function recommendation is defined to generate a list of the top 10 movie recommendations for a user.

The relatively high error indicated the need for exploring other methods to improve prediction accuracy, and because of that reason, additional models were considered for further study.

Non-negative Matrix Factorization (NMF)

The NMF model, operates by factorizing the user-item matrix into two lower-dimensional matrices. This factorization process uncovers latent factors that play a pivotal role in influencing user ratings. By breaking down the matrix, NMF is better suited for sparse data, can identify deeper and underlying patterns in user rating behaviors, and relationships in the data that are not immediately apparent.

For the implementation of the NMF model, the dataset was divided into two parts: a training set and a testing set. This division allowed for a comprehensive evaluation of the model's performance. The training set was used to train the model, enabling it to learn and adapt to the intricacies of the user-item interactions. The testing set, consisting of data not seen by the model during training, was then used to evaluate the model's effectiveness in predicting ratings on new, unseen data.

In this implementation, the NMF model achieved an MSE of 0.84. This significantly lower MSE, compared to the earlier Cosine Similarities model, indicated a considerable improvement in prediction accuracy.

Furthermore, this function identifies the top 10 movies for a particular user based on the model's predictions.

Singular Value Decomposition (SVD)

The SVD model, is a sophisticated matrix factorization technique. It decomposes the user-item rating matrix into singular vectors, capturing latent structures in the data, such as underlying user preferences and item characteristics. It's particularly effective in reducing the dimensionality of the data.

For the implementation, the dataset was split into training and testing sets to facilitate an unbiased evaluation of the model's predictive accuracy. The training set enables the model to learn from user-item interaction patterns, while the testing set, comprising data unseen during training, assesses the model's effectiveness on new information.

This model achieved an MSE of 0.75. This result indicates a higher accuracy in predicting user ratings compared to the earlier models, making it the most effective among the methods evaluated in this project.

As the other models, a list of the top 10 movies was generated at the end.

Literature

The MovieLens dataset, has been a cornerstone in the field of recommender systems. This section reviews recent academic contributions that have utilized this dataset.

Key Studies and Findings

Partitioned Models in Recommender Systems

In "Less Can Be More: Exploring Population Rating Dispositions with Partitioned Models in Recommender Systems" (2023), researchers investigated the effectiveness of segmenting users by their rating dispositions in recommender systems. The study's findings revealed that such partitioning not only improves computational efficiency but also enhances top-k performance and predictive accuracy, particularly in user-based KNN Collaborative Filtering systems. This indicates that adapting recommender algorithms to distinct user rating patterns can

significantly optimize both the system's processing capabilities and its ability to accurately predict user preferences.

Economics of Recommender Systems

The study "The Economics of Recommender Systems: Evidence from a Field Experiment on MovieLens" (2022), researchers focused on understanding the influence of recommendations on consumer behavior within the MovieLens platform. Their conclusions emphasized the profound informational impact of recommendations, going beyond mere exposure to new products. The study underscored that recommendations actively shape consumer beliefs and drive consumption choices. This points to the pivotal role of recommender systems in online marketplaces, not only in guiding consumers to products but also in forming their opinions and decision-making processes.

Connection between Projects

These studies provide valuable context for the current project. The first study's emphasis on adapting recommender systems based on user rating dispositions reflects the use of various algorithms (Cosine Similarities, NMF, SVD) to address diverse user preferences. Similarly, the second study's revelations about the powerful informational role of recommendations in shaping consumer decisions highlight the significance of developing precise predictive models, a central aim of our project. These studies affirm the importance of our approach towards enhancing the accuracy and quality of movie recommendations.

Building upon the insights from these studies, when compared with the first study, this project achieved an RMSE of 0.86 converted from the previously presented MSE, showcases commendable predictive accuracy. Although, using the same model (SVD), this is slightly higher than the study's RMSE of 0.75. This comparison indicates that while the project aligns well with the advanced methodologies used in current recommender system research, there is room for further optimization and fine-tuning to enhance the model's precision and narrow this gap in future iterations.

Despite the RMSE being slightly higher than the referenced study, the project still yielded positive results. This demonstrates the effectiveness of the methodologies used and shows promise for further enhancements in the realm of recommender systems.

Results and Conclusions

In assessing the performance of the models, the Cosine Similarities model served as a baseline but demonstrated a higher mean squared error (MSE) of 12.24, indicating lower prediction accuracy.

Because of that I had to study other options, and the second option was to use Non-negative Matrix Factorization (NMF) model, which improved upon this with a lower MSE of 0.83, capturing key latent features effectively.

However, it was the Singular Value Decomposition (SVD) model that outperformed the others, achieving the lowest MSE of 0.75. This superior performance of the SVD model is attributed to its efficient handling of sparse data and its ability to discern complex patterns in user-item interactions.

The notable differences in MSE values across the models stem from their varied technical methods in analyzing user-item data. Cosine Similarities, focusing on the angle between vectors, often fall short in sparse datasets due to their surface-level analysis. NMF, by factorizing the user-item matrix, delves deeper into latent features, offering better handling of sparse data. SVD, with its three-matrix decomposition, captures even more nuanced latent features and excels in reducing data dimensionality. This hierarchy in complexity and depth of analysis underlies the variation in predictive accuracy among these models.

To get prediction for a user, the user with the ID 1 was selected, and both models NMF and SVD models suggested some overlapping movies, given that the MSE is

close to each other, but given that the SVD is lower, its recommendations were deemed more reliable.

The top 10 recommendations for User 1, and the ratings from the SVD model are:

- The Great Escape (1963) - Predicted Rating: 5.0
- The Shawshank Redemption (1994) - Predicted Rating: 5.0
- Cool Hand Luke (1967) - Predicted Rating: 5.0
- Dr. Strangelove or: How I Learned to Stop Worrying and Love the Bomb (1964) - Predicted Rating: 5.0
- A Grand Day Out with Wallace and Gromit (1989) - Predicted Rating: 5.0
- A Streetcar Named Desire (1951) - Predicted Rating: 5.0
- Lawrence of Arabia (1962) - Predicted Rating: 5.0
- The Departed (2006) - Predicted Rating: 4.99
- Fight Club (1999) - Predicted Rating: 4.98
- The Dark Knight (2008) - Predicted Rating: 4.97.

References

Harper, F. M., & Konstan, J. A. (2015). The MovieLens Datasets: History and Context. ACM Transactions on Interactive Intelligent Systems (TiiS), 5(4), 19. Available at: https://grouplens.org/datasets/movielens/.

Sun, R., Kong, R., Jin, Q., & Konstan, J. (2023). Less Can Be More: Exploring Population Rating Dispositions with Partitioned Models in Recommender Systems. In Adjunct Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization (UMAP '23 Adjunct), 291–295. DOI: 10.1145/3563359.3597390.

Aridor, G., Goncalves, D., Kluver, D., Kong, R., & Konstan, J. (2022). The Economics of Recommender Systems: Evidence from a Field Experiment on MovieLens. DOI: 10.48550/arXiv.2211.14219.