A Movie Recommendation System Using MovieLens Data

Authors: Delvin Osoro, Enid Kibet, Logan Njiru, Winniefred Kinyumu, Dennis Limo

---

Executive Summary

This report outlines the development and analysis of a movie recommendation system based on the MovieLens dataset. The system addresses the challenge of choice overload faced by users on streaming platforms by offering personalized movie suggestions. By leveraging collaborative filtering and content-based filtering, the system delivers accurate and tailored recommendations. Through detailed data preparation, exploratory analysis, and advanced modeling techniques, the project achieved meaningful insights and reliable recommendations, creating a foundation for real-world applications.

Business Understanding

In today's streaming landscape, platforms like Netflix, HBO, and Showmax provide vast movie libraries. While variety attracts users, it often leads to decision fatigue, making it harder for users to find content aligned with their tastes. Personalized recommendations not only enhance user satisfaction but also improve engagement and retention.

The goal of this project was to build a system capable of recommending the top 5 movies most suited to a user's preferences. Using the MovieLens dataset, the project identified user preferences through ratings and movie attributes. By addressing challenges like data sparsity and understanding user behaviors, the system provides a scalable, effective solution for improving user experiences on streaming platforms.

Problem Statement

With thousands of movies available on streaming platforms, users often feel overwhelmed by the choices. This choice overload reduces satisfaction and engagement. To solve this problem, the project aimed to develop a personalized recommendation system that could predict and suggest movies tailored to individual users' preferences. This system enhances user experience by offering curated, relevant movie suggestions.

Project Objectives

1. Identify the top 10 most popular genres based on user ratings.
2. Analyze the top 10 genres with the highest average ratings.
3. Examine the distribution of movie ratings per user.
4. Identify trends in movie production by decade.
5. Assess the number of ratings per movie to address data sparsity.
6. Determine the most popular movies using weighted metrics like the Bayesian average.
7. Investigate the relationship between movie genres and user ratings.
8. Examine user-generated tags for identifying popular keywords associated with movie themes.
9. Develop a recommendation system to suggest the top 5 rated movies for a user using a hybrid approach.

Stakeholders

The key stakeholders for this project include:

- **Streaming Platform Users**: End-users who will benefit from personalized recommendations that enhance their viewing experience.
- **Content Providers**: Streaming platforms that can use the system to boost user retention and engagement.
- **Business Executives**: Decision-makers who will leverage insights from this system to drive strategic initiatives and improve platform performance.
- **Data Science Teams**: Professionals responsible for maintaining and enhancing the recommendation system to keep it relevant and effective.

Dataset Overview

The MovieLens dataset, provided by GroupLens research, is a rich source of movie ratings and metadata. The key components of the dataset include:

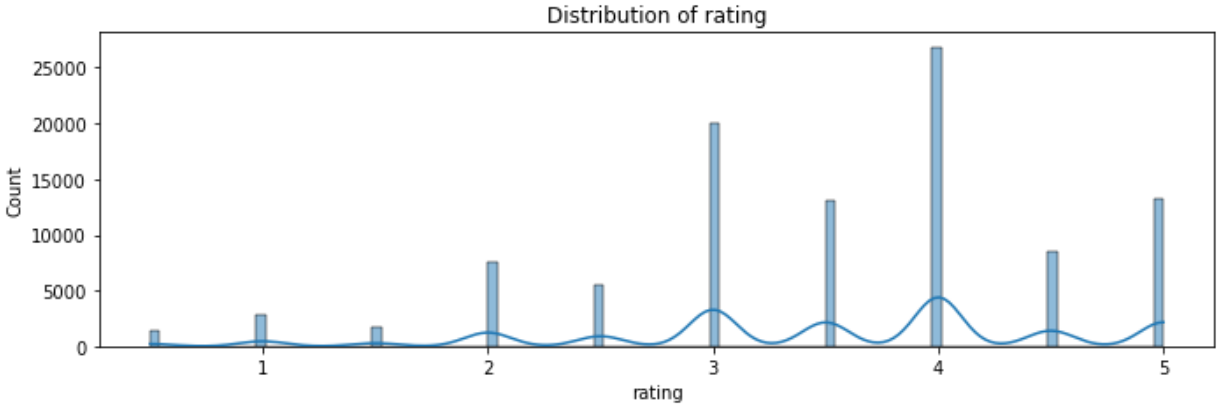
- **Movies.csv**:
    - `movieId`: Unique identifier for each movie.
    - `title`: The title of the movie along with its release year.
    - `genres`: List of genres associated with the movie.
- **Ratings.csv**:
    - `userId`: Unique identifier for each user.

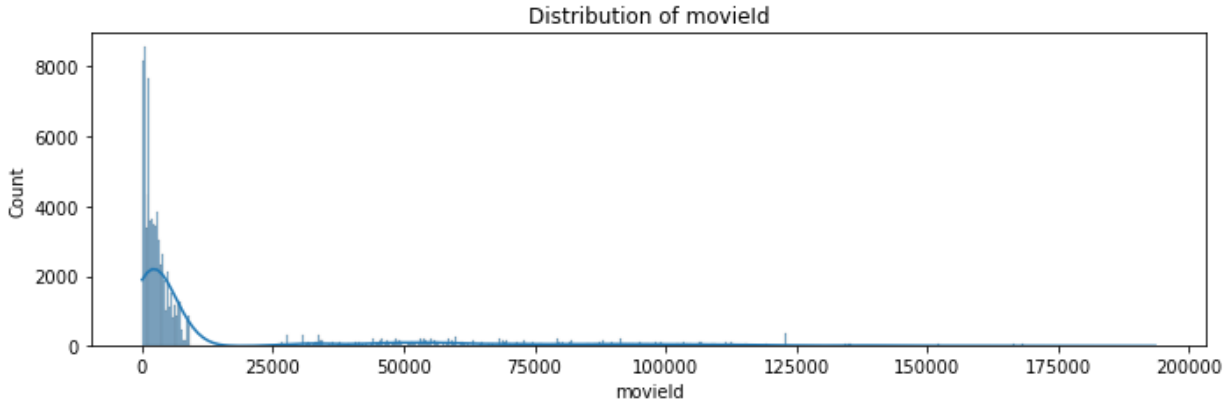
- `movieId`: Identifier linking the movie to the Movies.csv file.
- `rating`: User-provided rating on a scale of 0.5 to 5.0.
- `timestamp`: The time when the rating was provided.

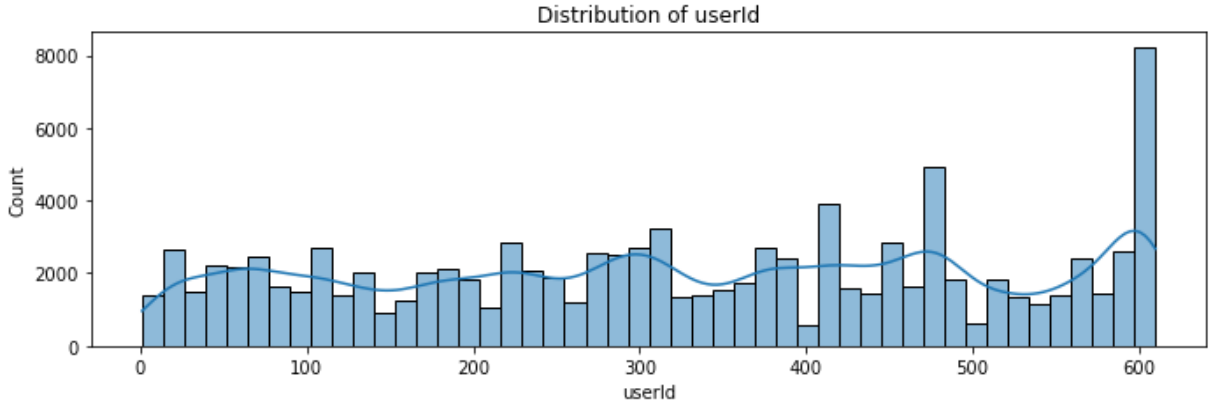
Data Processing and Cleaning

The data processing and cleaning phase ensured that the dataset was accurate, consistent, and ready for analysis. Here are the detailed steps:

1. **Accuracy**:
    - Checked for duplicate rows in each dataset. No duplicates were found, confirming the data's integrity.
2. **Validity**:
    - Converted the `timestamp` column in the `ratings` dataset into a human-readable date format. However, it was later dropped since all timestamps were set to the same default value, making it irrelevant to the analysis.
    - Extracted the release year from the `title` column in the `movies` dataset. This information was added as a separate column (`Release_year`) for temporal analysis.
3. **Completeness**:
    - Assessed the percentage of missing values in all datasets. None were found, confirming the data was complete and ready for use.
4. **Uniformity**:
    - Standardized column naming conventions to ensure consistency across datasets.
    - Exploded the `genres` column, which contained multiple genres separated by pipes (`|`), into individual rows. This step allowed for more granular analysis of each genre.
5. **Distribution Analysis**:
    - Examined the distribution of key numerical variables (`ratings`, `movieId`, and `userId`) to understand the data's shape and coverage. Most ratings were concentrated between 2.0 and 4.5, reflecting positive sentiment. However, a large number of movies had very few ratings, highlighting the sparse nature of the dataset. Understanding these distributions informed decisions on handling sparsity and extreme values during model development.

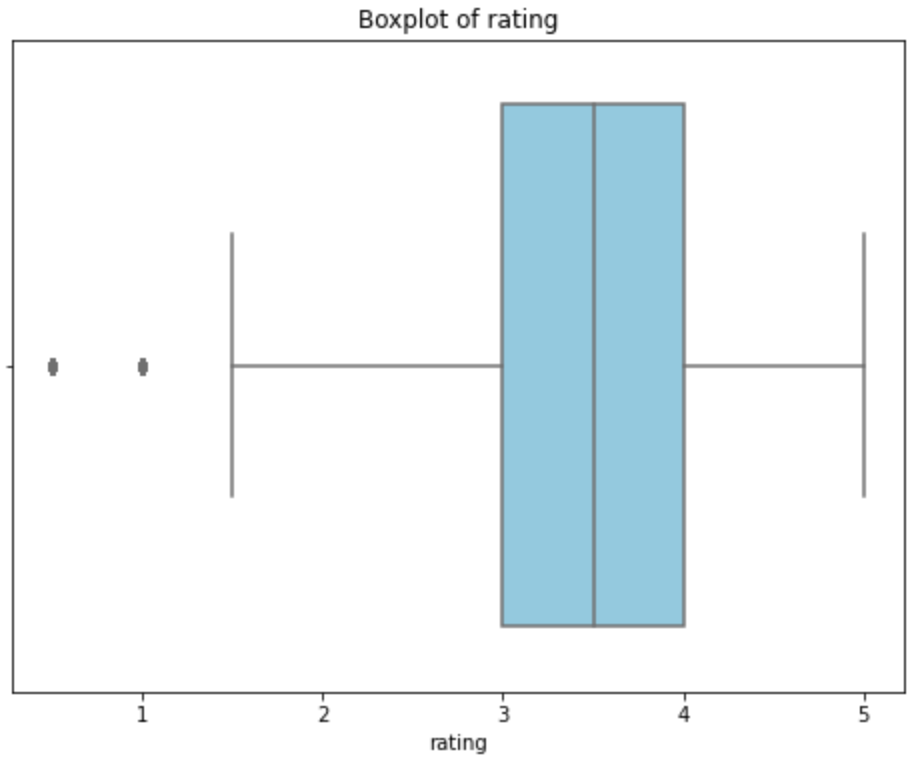






6. **Outliers**:

   - Outliers in user ratings were identified through boxplots, showing some very low and very high ratings. However, outliers were not removed because in the context of recommendation systems, extreme ratings provide critical insights into user preferences (e.g., strong likes or dislikes). Removing them would lead to a loss of valuable information necessary for personalized recommendations.

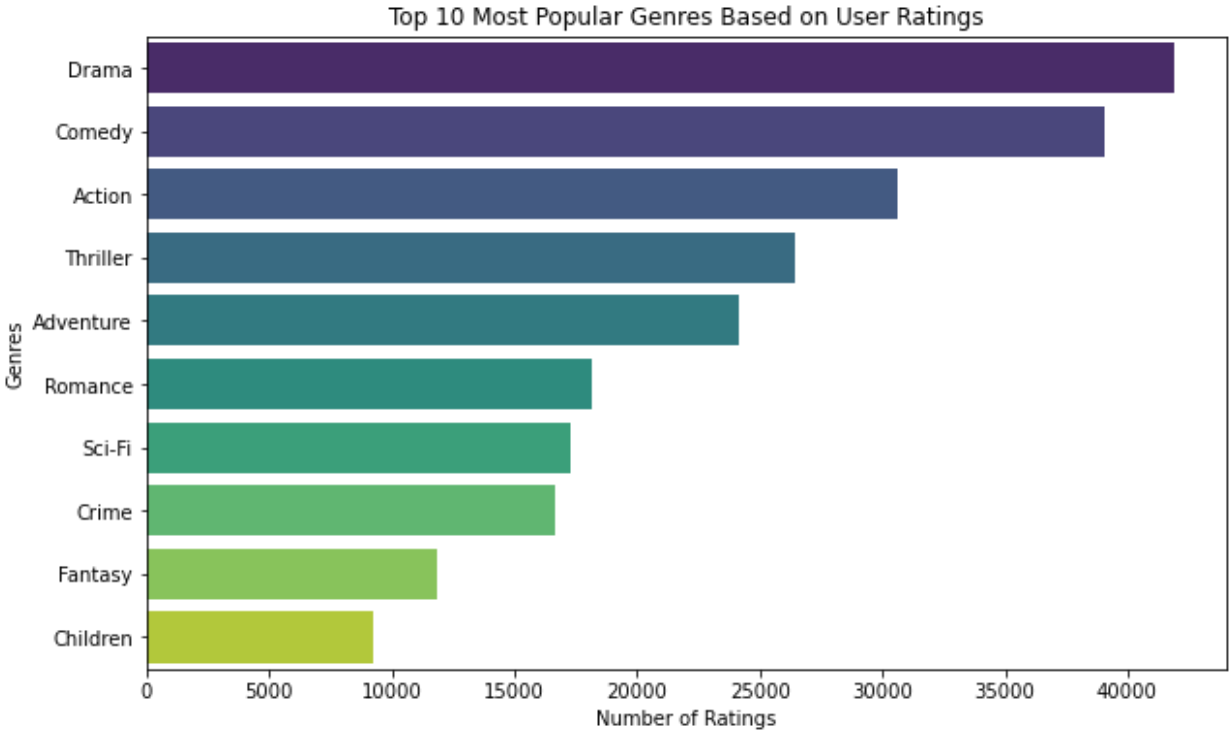


7. **Merging**:
    - Combined the `Movies.csv` and `Ratings.csv` datasets using the `movieId` column as the key. The resulting dataset provided a comprehensive view of movies, their attributes, and the associated user ratings. This merged dataset formed the foundation for the exploratory data analysis and modeling phases.

Exploratory Data Analysis (EDA)

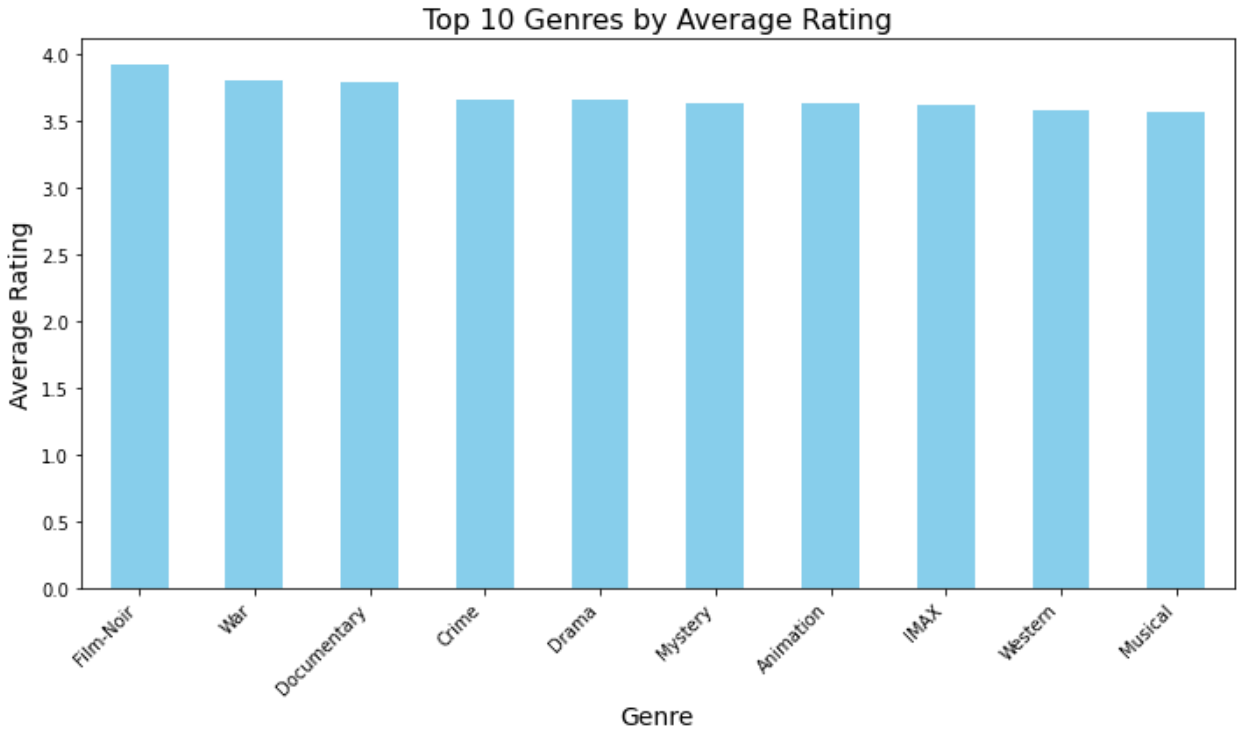
1. Most Popular Genres

The analysis revealed the top 10 most-rated genres, with **Drama** leading with 41,928 ratings, followed by **Comedy** with 39,053 ratings, and **Action** with 30,635 ratings. These genres represent the bulk of user interactions, indicating their popularity and relevance in guiding recommendations. The visualization of this data showed a clear dominance of these genres, emphasizing their central role in user engagement.



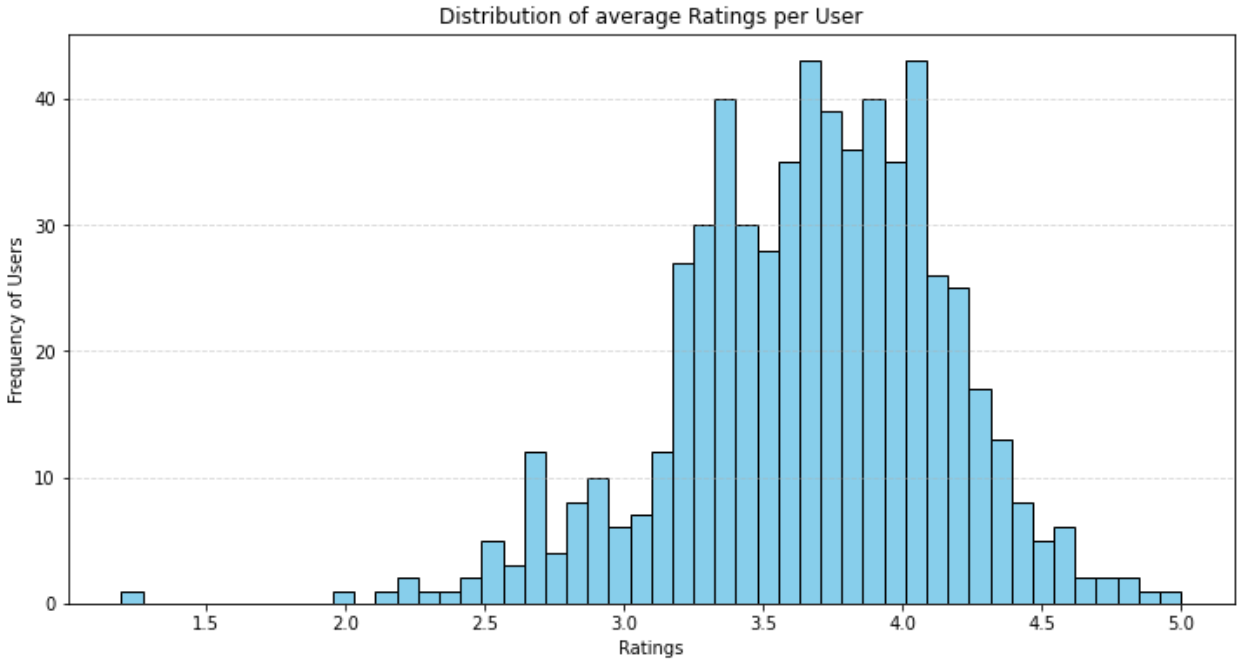
2. Genres with Highest Ratings

When analyzing average ratings, **Film-Noir**, **War**, and **Documentary** emerged as the highest-rated genres. While these genres may not have the highest number of ratings, their consistently high scores reflect strong appreciation from a niche audience. This insight is critical for catering to specific user groups with distinct preferences.



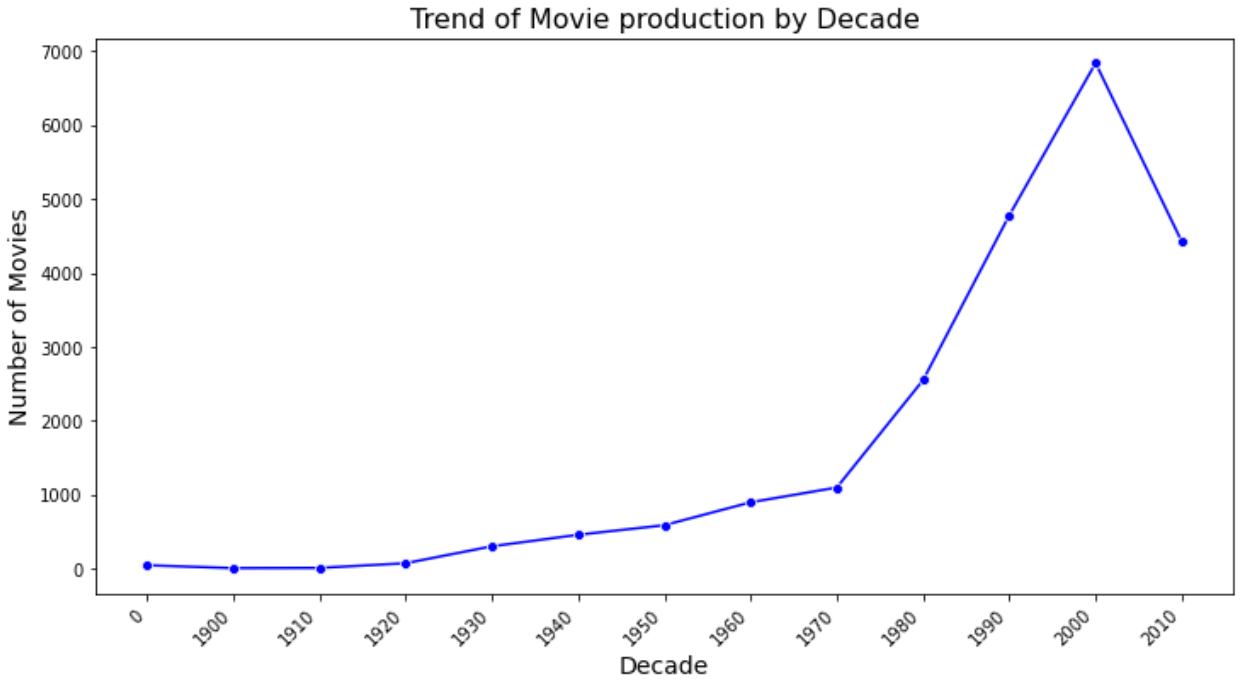
3. Distribution of Ratings Per User

The distribution of ratings revealed that most users tend to give ratings between 2.0 and 4.5, reflecting a generally positive sentiment toward movies. However, the presence of low ratings (below 2.0) indicates occasional dissatisfaction, which highlights the need for accurate recommendations to minimize mismatches.



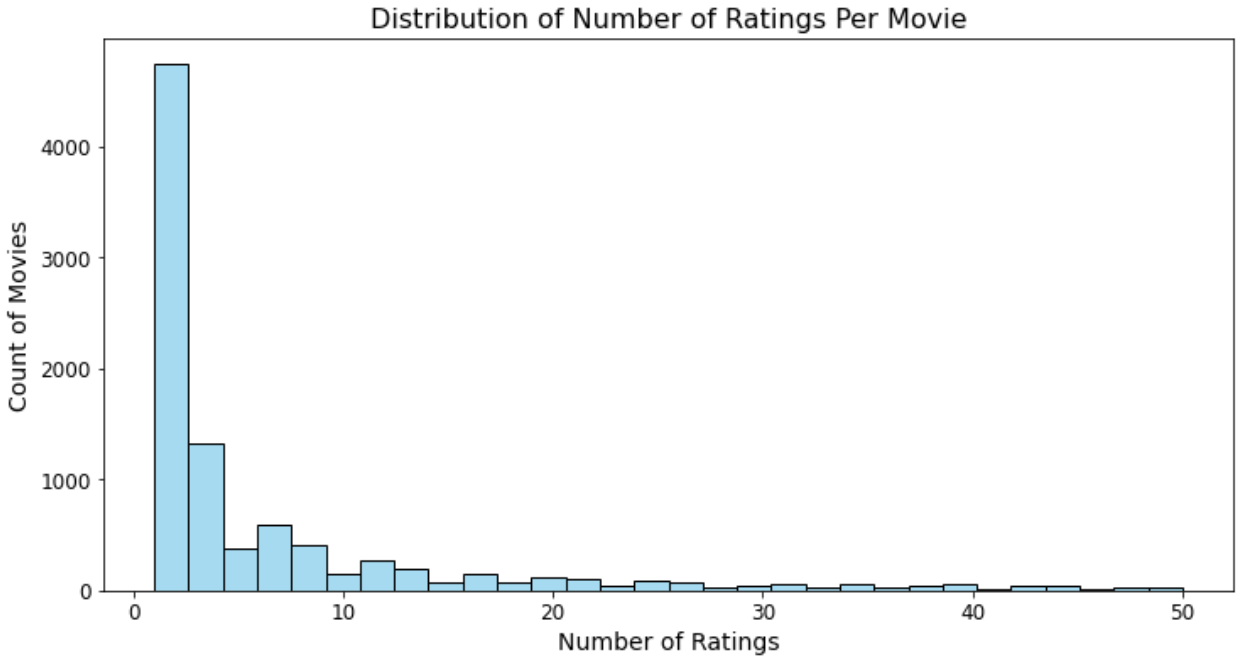
4. Trends in Movie Production by Decade

Analyzing the release years of movies showed that production significantly increased from the 1980s to the 2000s, peaking during this period. This growth aligns with advancements in technology and increased global demand for entertainment. However, the slight decline in the 2010s could be attributed to the rise of alternative content formats, such as web series and short films. This trend provides an opportunity to highlight classic movies from high-production decades.



5. Number of Ratings Per Movie

The majority of movies received fewer than 10 ratings, revealing a sparse distribution. However, a smaller subset of movies garnered a significant number of ratings, indicating their widespread appeal. This underscores the importance of addressing data sparsity in recommendations and ensuring less-rated movies are not overlooked.

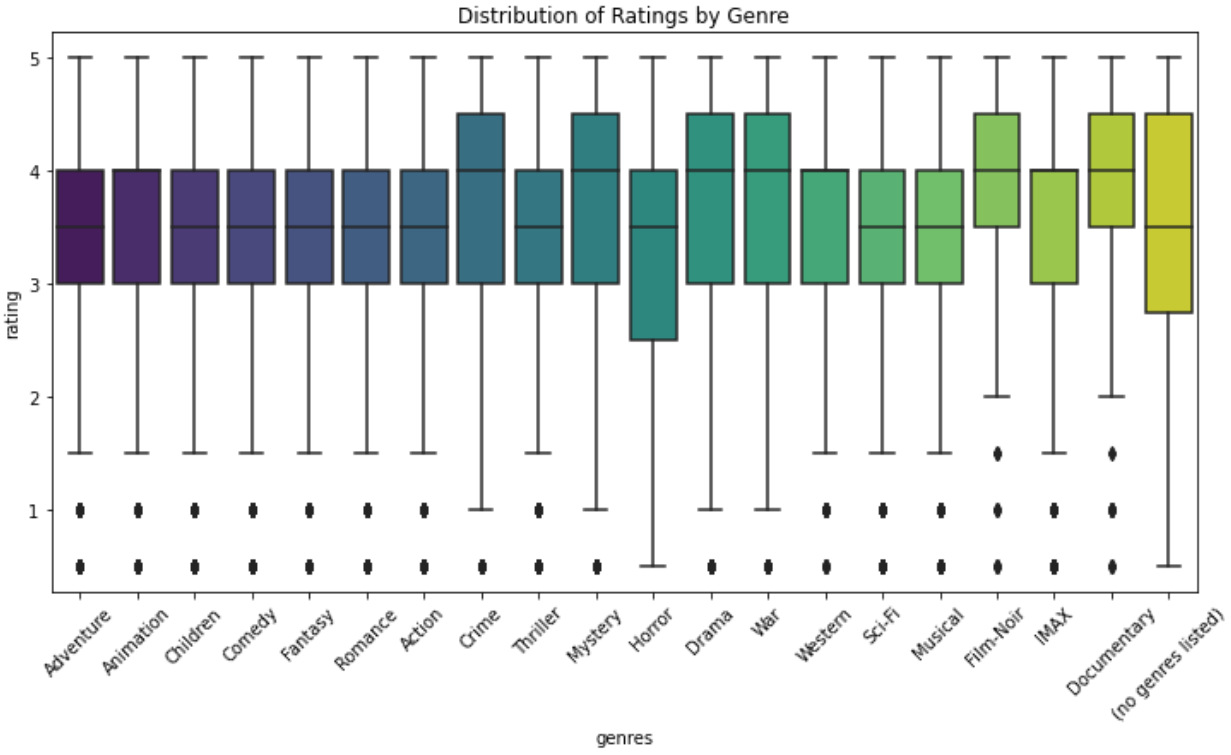


6. Most Popular Movies (Weighted)

Using the Bayesian average, which balances the number of ratings with their average score, provided a more reliable measure of a movie's popularity. This method helped identify movies that are both widely rated and highly appreciated, ensuring fairness in ranking across the dataset.

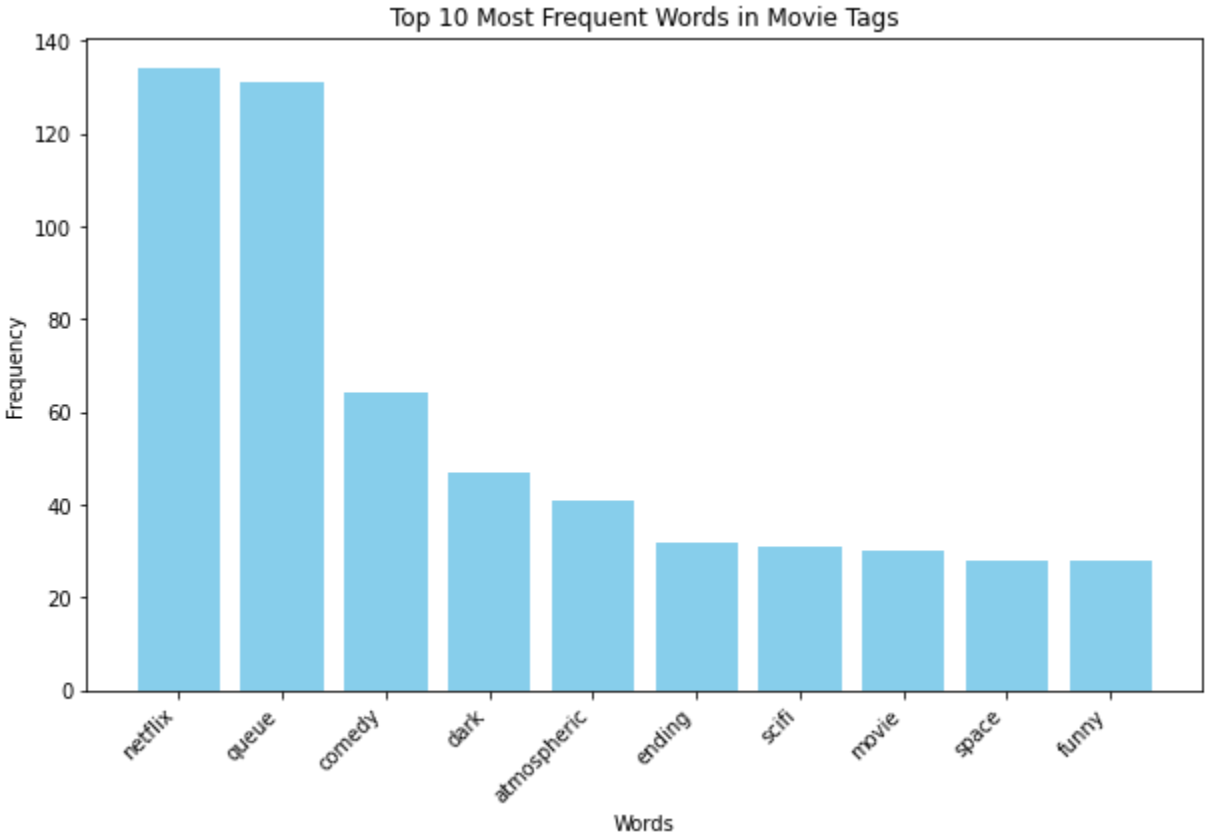
7. Relationship Between Genres and Ratings

A statistical analysis (one-way ANOVA) confirmed a significant relationship between genres and ratings, with some genres consistently receiving higher ratings than others. This finding supports the idea that user preferences are heavily influenced by genre, which is a critical factor in designing tailored recommendations.



8. User-Generated Tags

Though not explored in detail, future work could examine tags provided by users. These tags often include keywords such as "funny," "dark," or "sci-fi," which could be leveraged to enhance content-based recommendations and provide additional context for movie preferences.



Model Development

1. Collaborative Filtering

The Singular Value Decomposition (SVD) algorithm was implemented to predict user ratings based on patterns in the user-item interaction matrix. SVD achieved the lowest RMSE (0.8501), outperforming other models like KNNBasic and KNNBaseline. SVD works by breaking down the sparse matrix of user-item interactions into smaller, dense matrices that capture hidden patterns, enabling the prediction of missing ratings.

Metric of Success: The success metric for the SVD model was the Root Mean Squared Error (RMSE), which evaluates the difference between predicted and actual ratings. The model achieved an RMSE of 0.8501, indicating accurate predictions.

2. Content-Based Filtering

Content-based filtering was used to recommend movies by analyzing their attributes, such as genres. The algorithm matches user preferences (e.g., movies they have highly rated) with other movies that share similar characteristics. For instance, if a user frequently rates movies in

the "Drama" genre highly, the system suggests other highly-rated drama films. This approach complements collaborative filtering by focusing on the properties of items rather than user interactions.

3. Hybrid Approach

The hybrid approach combines the strengths of collaborative filtering and content-based filtering. It utilizes SVD to predict ratings based on user-item interactions while simultaneously incorporating movie attributes like genres. This integration helps address challenges like the cold start problem, enabling the system to recommend movies even for new users or less-rated items. The hybrid model ensures a more personalized and diverse recommendation experience.

Metric of Success: The hybrid approach was evaluated based on its ability to address cold start challenges while maintaining accurate predictions. It successfully bridged gaps where collaborative filtering alone could not provide recommendations.

Recommendations

1. **Leverage Ratings to Build Personalization**
   Encourage users to rate more movies, as the system effectively uses these ratings to recommend movies. Additional data can improve the accuracy and relevance of recommendations by providing richer insights into user preferences.

2. **Focus on a Hybrid Approach**
   The hybrid recommendation system, which combines collaborative filtering with content-based filtering, successfully addresses challenges like the cold start problem. This dual approach ensures a more tailored user experience by considering both user preferences and movie attributes.

3. **Enhance Genre-Based Suggestions**
   Since genres significantly influence user preferences, consider building recommendations that prioritize a user’s favorite genres. Providing curated lists based on highly rated genres can enhance engagement and satisfaction.

4. **Adapt to Shifting Trends in Movie Production**
   Observing production trends across decades highlights user interest in specific eras. Streaming platforms can leverage these insights to showcase popular movies from specific time periods, appealing to nostalgia or curiosity about classic films.

5. **Encourage Ratings from New Users**
   Simplify the onboarding process by allowing new users to rate a few movies. This initial engagement reduces cold start issues and enables the system to provide accurate recommendations early in the user experience.

6. **Explore Metadata Beyond Ratings**
   Future iterations could incorporate metadata such as cast, director, or production year to refine content-based filtering further. These additional attributes can help identify nuanced patterns in user preferences, enabling more personalized recommendations.

By implementing these strategies, the recommendation system can remain user-focused, foster engagement, and cater to diverse audience preferences with meaningful and personalized suggestions.

Conclusion

This project successfully developed a robust movie recommendation system using the MovieLens dataset. By combining collaborative and content-based filtering, the system provided accurate and personalized suggestions, addressing common challenges such as data sparsity and the cold start problem. Future deployment and continuous refinement of the system will ensure scalability, user satisfaction, and long-term success.