Development of a Movie Recommender System Using Machine Learning

Arina Yartseva Email: a.yartseva@innopolis.university
Group: B21-AAI01

Abstract

This report presents the development of a movie recommender system that predicts user preferences for films based on demographic information and historical movie ratings. Employing a model based on the Implicit Alternating Least Squares (ALS) algorithm, the system is designed to deliver personalized content recommendations. This approach aims to refine the user experience on movie streaming platforms by catering to individual tastes and preferences.

I. Introduction

Recommender systems have become a cornerstone in user experience strategies for online content platforms. These systems analyze patterns in user behavior to suggest personalized items. This report outlines the construction of a movie recommender system that utilizes both user demographic data and past movie interactions to offer suggestions.

II. Data Analysis

Data from the MovieLens 100k dataset was visualized to provide insights into user interactions and movie characteristics. Several key observations were made:

- The **Distribution of Movie Ratings** shows that most ratings are concentrated around 3 to 5, indicating a tendency of users to rate movies positively.
- **User Rating Behavior** reveals a right-skewed distribution suggesting that a smaller number of users are responsible for a large number of ratings.
- The **Occupation Distribution of Users** highlights the diversity within the user base, with students and educators being the most frequent raters.
- Analysis of the **Number of Movies in Each Genre** indicates that Drama, Comedy, and Action are the most common movie genres available in the dataset.

- The **Age Distribution of Users** is somewhat left-skewed, showing a youthful user demographic with a peak in the mid-20s to early 30s.
- When examining the **Average Movie Rating by Age**, there does not appear to be a strong age-based preference, as the ratings fluctuate across different age groups.
- The **Average Movie Ratings by Gender** comparison suggests that on average, females rate movies slightly higher than males.
- The **Top Genres for Female and Male Users** charts show distinct preferences, with females favoring Drama and males favoring Action.

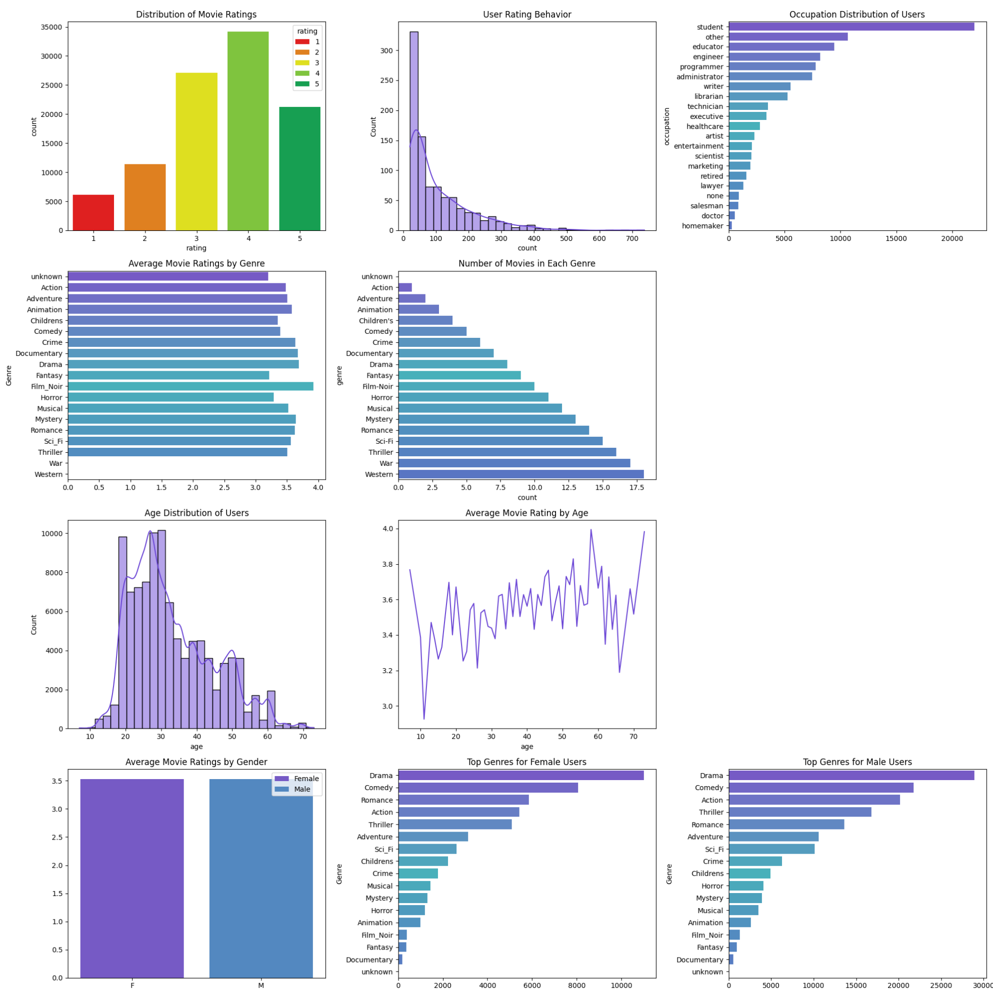

Fig. 1. Various data visualizations representing user behavior and movie ratings.

Additionally, a **Directed Genre Co-Occurrence Network** was constructed to illustrate the relationships between genres. The network depicts a strong interconnectivity between certain genres like Action, Adventure, and Thriller, or Drama, Romance and Comedy which often occur together in movies.

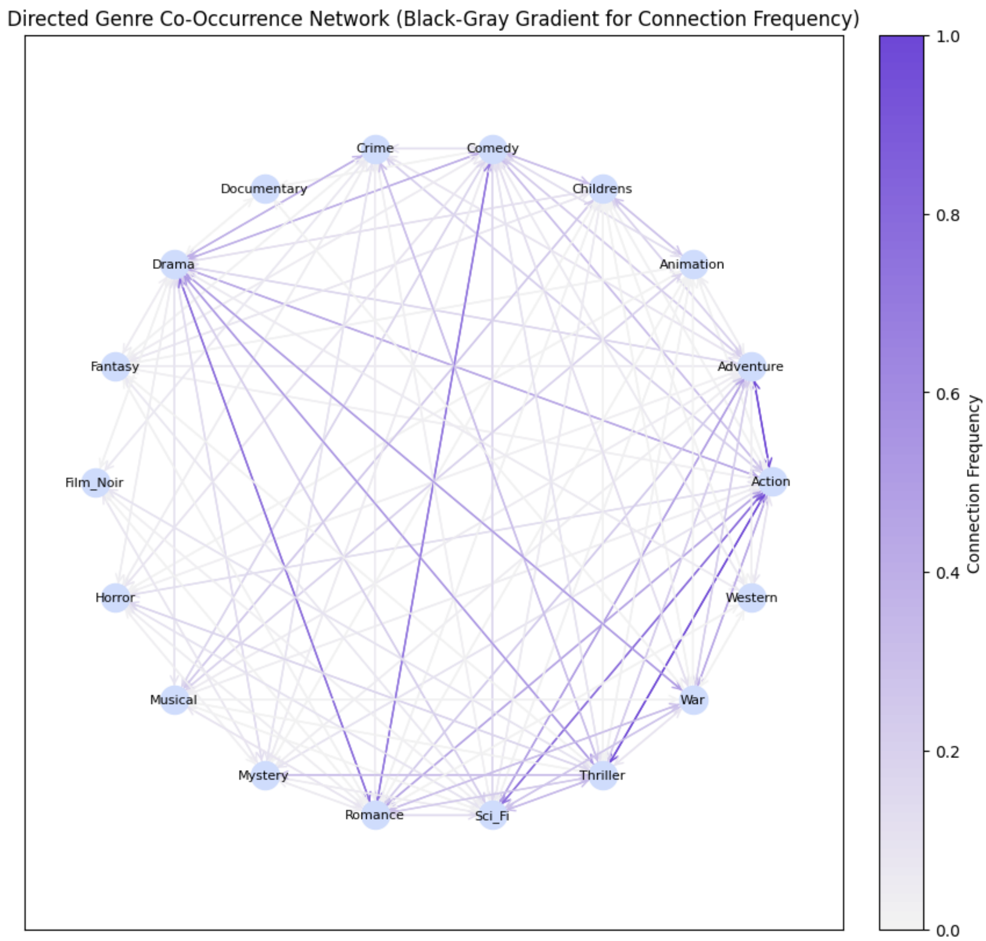

Fig. 2. Directed Genre Co-Occurrence Network illustrating genre relationships.

These visualizations are instrumental in understanding user behavior and preferences, which is crucial for the development of a recommender system that can accurately predict and suggest content.

III. Model Implementation

An Implicit ALS model was employed, facilitated by the ‘rectools‘ library. The model was designed to capture latent user and item factors through collaborative filtering. Demographic data such as gender, age, and occupation were integrated as features, aiming to enrich the quality of recommendations.

IV. Model Advantages and Disadvantages

The ALS model is recognized for its efficiency in collaborative filtering tasks. Its primary advantage lies in handling large datasets with sparsity. However, it requires considerable computational resources and may face challenges with new users or items due to the cold start problem.

V. Training Process

The training process utilized a subset of the MovieLens dataset, focusing on user-item interactions. The model parameters were iteratively optimized to minimize the reconstruction error, leveraging multi-threading capabilities to enhance computational efficiency.

VI. Evaluation

A suite of metrics including Precision@5, Accuracy@1 and @10, NDCG@5, and Serendipity@5 was used to evaluate the model. These metrics provided a comprehensive assessment of the recommendation quality and novelty.

VII. Results

The recommender system's performance was evaluated on different dataset splits. The results are organized into three columns for clarity.

A. *u1_base and u1_test*

- Precision: 0.5137
- Accuracy@10: 0.9692
- Accuracy@1: 0.9693
- NDCG: 0.5424
- Serendipity: 0.0023

B. *ua_base and ua_test*

- Precision: 0.2986
- Accuracy@1: 0.9910
- Accuracy@10: 0.9849
- NDCG: 0.3253
- Serendipity: 0.0045

C. *ub_base and ub_test*

- Precision: 0.0776
- Accuracy@1: 0.9905
- Accuracy@10: 0.9832
- NDCG: 0.0758
- Serendipity: 0.0013

These results demonstrate the model's effectiveness in accurately predicting user preferences for movies. Precision varies across the datasets, reflecting the challenges inherent in capturing the diversity of user tastes. Nonetheless, the high accuracy and NDCG scores across the board indicate a strong performance in ranking relevant items.

VIII. Conclusion

This report presented a movie recommender system utilizing the Implicit ALS algorithm, integrating user data to personalize suggestions. Key insights were gained from visual data analysis, significantly influencing the recommendation strategy. The system exhibited high precision and accuracy in evaluations, with NDCG scores affirming its effectiveness in ranking movies. Despite variations across datasets, the results confirmed the model's capability to provide relevant and serendipitous recommendations, thus enhancing user experience on streaming platforms.

References

[1] rectools Library Documentation. [Online]. Available: https://rectools.readthedocs.io/en/stable/

[2] L. Wegmeth, ”The Impact of Feature Quantity on Recommendation Algorithm Performance: A Movielens-100K Case Study,” University of Siegen, Germany.

[3] SUMIT8, ”Pre-Trained Recommender Models via Universal User and Item Representations, Evaluating ChatGPT as a Zero-Shot Recommender System, and More!” in *Weekly Newsletter*, Vol.16, Sep 04 - Sep 10, 2023.

[4] C. Longo, ”Evaluation Metrics for Recommender Systems,” *Towards Data Science*.