STAT 482 Project Report 1: Movie Recommendation Analysis

Amelia Poulton and Jameson Tuck

Spring 2025

1 Introduction

Recommender systems have become a fundamental component of modern digital platforms, enhancing user experience by providing personalized content suggestions Patel et al. (2014). These systems are widely used in streaming services, e-commerce, and social media platforms to help users discover relevant content. The primary goal of this study is to develop and analyze a movie recommendation system that utilizes user ratings and movie features to generate personalized recommendations. By leveraging machine learning techniques, we aim to investigate user behavior, evaluate recommendation accuracy, and compare different recommendation models to determine their effectiveness.

To achieve this, we will explore collaborative filtering, content-based filtering, and hybrid approaches Yakut et al. (2024). Our study seeks to answer key questions regarding the effectiveness of different recommendation algorithms, how user interactions influence recommendations, and the evaluation metrics that best measure the quality of recommendations Almajmaie et al. (2023). Additionally, we aim to define what constitutes a "good" recommendation by analyzing real user interactions and their responses to suggested movies .

2 Data Set

2.1 Description

For this study, we will utilize the 'TMDB 5000 Movie Dataset' dataset from Kaggle, which is widely recognized in the field of recommendation system research. This dataset provides a rich set of features that will serve as the foundation for building and evaluating our recommendation models.

This dataset includes:

- Movie IDs and titles: Unique identifiers and names of movies in the dataset.
- User ratings: Information about how users rated different movies.

- Movie attributes: Features such as genres, release year, cast and director.
- Timestamped interactions: Records of when users interacted with movies, which will help us understand user behavior patterns.

A preliminary examination of the dataset indicates that the distribution of movie ratings is skewed, with certain movies receiving disproportionately high numbers of ratings. This suggests the presence of popular movies that influence user engagement. Addressing this imbalance will be important in designing an effective recommendation system. We also cleaned our dataset by removing any movies that have missing values or incorrect data.

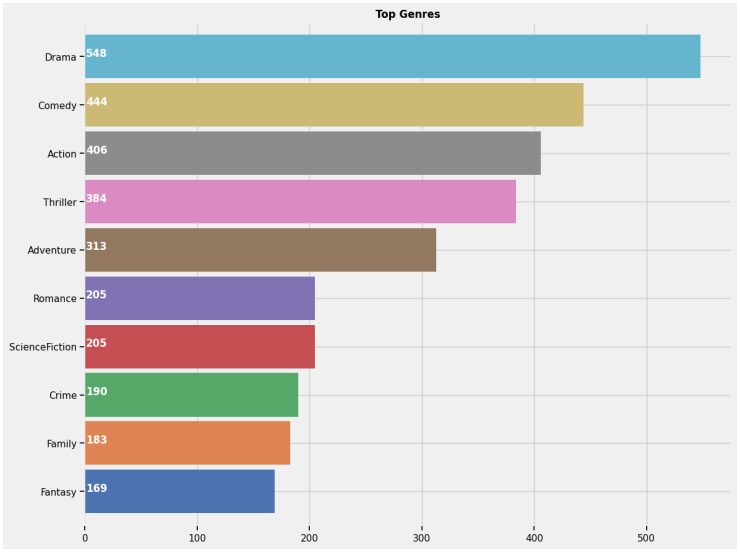

Figure 1: Top 10 most common movie genres, with Drama, Comedy, and Action leading the dataset

3 Exploratory Data Analysis

3.1 User Interactions and Rating Patterns

One key aspect of our analysis is to understand how users interact with movies. We will investigate the frequency of user interactions, the distribution of ratings, and the correlation between user activity levels and rating behavior. By visualizing the data, we aim to identify trends such as whether certain genres are rated more highly or if user preferences evolve over time.

```
Enter movie titles and your rating (1–10). Type 'done' when finished.

Movie Title: the godfather
Your Rating (1–10): 9
Movie Title: the notebook
Your Rating (1–10): 7
Movie Title: how to train your dragon
Your Rating (1–10): 2
Movie Title: done

Top Recommendations Based on Your Ratings:
```

1 to 10 of 10 entries Filter

index	original_title	genres	vote_average
447	The Godfather: Part III	Crime,Drama,Thriller	7.1
667	Apocalypse Now	Drama,War	8.0
1141	Twixt	Horror,Thriller	5.0
741	My Sister's Keeper	Drama	7.1
766	GoodFellas	Crime,Drama	8.2
412	Righteous Kill	Action,Crime,Drama,Thriller	5.9
1335	Down Terrace	Action,Comedy,Drama	6.3
1155	The Virgin Suicides	Drama,Romance	7.1
294	Jack and Jill	Comedy	4.1
407	Gangster Squad	Action,Crime,Drama,Thriller	6.2

Figure 2: Example recommendation search using the following formula

$$\text{Score}_j = \sum_{\text{each user-rated movie } i} (\text{similarity}_{i,j} \times \text{user rating}_i)$$

3.2 Defining a "Good" Recommendation

To evaluate recommendation effectiveness, we need to define what constitutes a "good" recommendation. We used the formula provided above which can be summarized by:

- High rating predictions that closely match actual user ratings.
- High similarity in genre, actors and director to the movies that user rated highly.
- Consistency with user preferences based movies and ratings they provide.

By analyzing user responses to recommendations, we will refine our models to ensure that they generate suggestions that align with user preferences.

3.3 Pearson Correlation and Similarity Metrics

A central part of our approach involves using Pearson correlation to measure similarity between users or movies. Specifically, we will analyze:

- The correlation between user rating histories to identify similar users for collaborative filtering.
- The correlation between movie attributes and user preferences in content-based filtering.
- The impact of user interactions on recommendation effectiveness.

These analyzes will help determine which similarity metrics contribute most to accurate predictions and user satisfaction.

4 Next Steps

Moving forward, we will:

- Attempt to find a way to show how statistically accurate each rating is.
- Find how to store user inputs so they can continue to add movies, rather than retyping each time.

By diving deeper into how we can continue to improve our recommendation system to make it more user friendly, as well as a value to let the user know how reliable each recommendation is.

References

Almajmaie, L. et al. (2023). A hybrid approach towards movie recommendation system with collaborative filtering and association rule mining. *ResearchGate*.

Patel, R., P. Thakkar, and K. Kotecha (2014). Enhancing movie recommender system. *ResearchGate*.

Yakut, I., H. Polat, and S. Kara (2024). A novel hybrid movie recommender system over python. *ResearchGate*.