DSCI 632 - Applied Cloud Computing

Building a Movie Recommendation Engine and Forecasting User Ratings

Aman Ostwal (ago34@drexel.edu)
Darshit Rai (dr3264@drexel.edu)
Sanskruti Chavanke (sc4323@drexel.edu)

11th March, 2024

Contents

List of Figures

Abstract

Recommender systems play a pivotal role in the functionality of various digital platforms like Netflix, Amazon, and others, contributing significantly to user engagement. This study delves into the examination and comparison of two methodologies for recommendation systems: the ***Content-Based Filtering*** approach employing the ***Jaccard Index*** and the ***Collaborative Filtering Method*** utilizing ***Alternating Least Squares (ALS)*** matrix factorization. While the Jaccard Index approach demonstrated anticipated performance, it encountered challenges related to insufficient data. In contrast, the ALS method proved effective in predicting user content ratings, achieving a Root Mean Square Error (RMSE) of **0.912023**.

1 Introduction

Recommender systems have developed into essential resources for maintaining and improving user interaction on a variety of digital service platforms, including Netflix, Hulu, and Amazon. Different methodologies are used in the field of content recommendation; these include hybrid approaches, session-based systems, reinforcement learning systems, risk-aware systems, collaborative filtering, and content-based filtering.

In this paper, the mechanisms and effectiveness of two different recommendation methodologies—the content-based filtering Jaccard Index method and the collaborative filtering Alternating Least Squares (ALS) matrix factorization approach—are thoroughly explored, compared, and contrasted. GroupLens has donated the MovieLens dataset for the study, which provides insightful information about the functionality and complexities of these recommendation systems.

2 Challenges and Ethical Consideration

The information we collected might unintentionally lean towards the preferences of certain groups, making our recommendations somewhat biased. This could affect how fair our system is for everyone. Sometimes, there's not enough information about user ratings, especially for less-known movies. This might make our system less accurate, particularly when suggesting these kinds of films. It's super important that we let users know we're collecting and using their data to recommend movies. Getting their permission is key to making sure everything is above board and respectful of their privacy. We aim to be clear about how our recommendation system works. This way, users can understand why they're getting certain suggestions, making the whole process more trustworthy.

3 MovieLens Dataset Overview

The MovieLens dataset, sourced from GroupLens, encompasses around **25000095** ratings spanning approximately **59047** movies and involving roughly **162541** users. These ratings were contributed between January 9th, 1995, and November 21st, 2019. This dataset was freely obtained from grouplens.org. Each entry in the dataset includes the movie ID, the user ID assigning the rating, and a timestamp representing the rating time in Unix seconds since 1/9/1995. Supplementary tables are provided to facilitate linking and incorporating movie titles into the dataset. The focal variable is the rating, ranging from 0 to 5. Table 1 presents a breakdown of the sample variables within this dataset.

```
Number of Ratings: 25000095
Number of Distinct Users: 162541
Number of Distinct Movies: 59047
```

Figure 1: MovieLens Dataset Overview

4 Exploratory Data Analysis (EDA)

4.1 Analysis of Movie Popularity

In Figure 2 and Figure 3, we present graphical representations depicting the ratings of the most and least popular movies within the dataset. Notably, the highest-rated films, led by the iconic "Forrest Gump" released in 1994, are followed closely by other classics such as "Shawshank Redemption" and "Pulp Fiction." These movies have consistently garnered favourable ratings, establishing them as the most popular among viewers.

Conversely, Figure 3 sheds light on the least popular movies, struggling to secure favourable ratings, often receiving only 1 out of 5. Examples of such less-favoured films include "The Olsen Gang in a Fix" from 1969, "Confessions from a Holiday," and "The Most Beautiful Wife," suggesting a discernible contrast with the higher-rated counterparts.

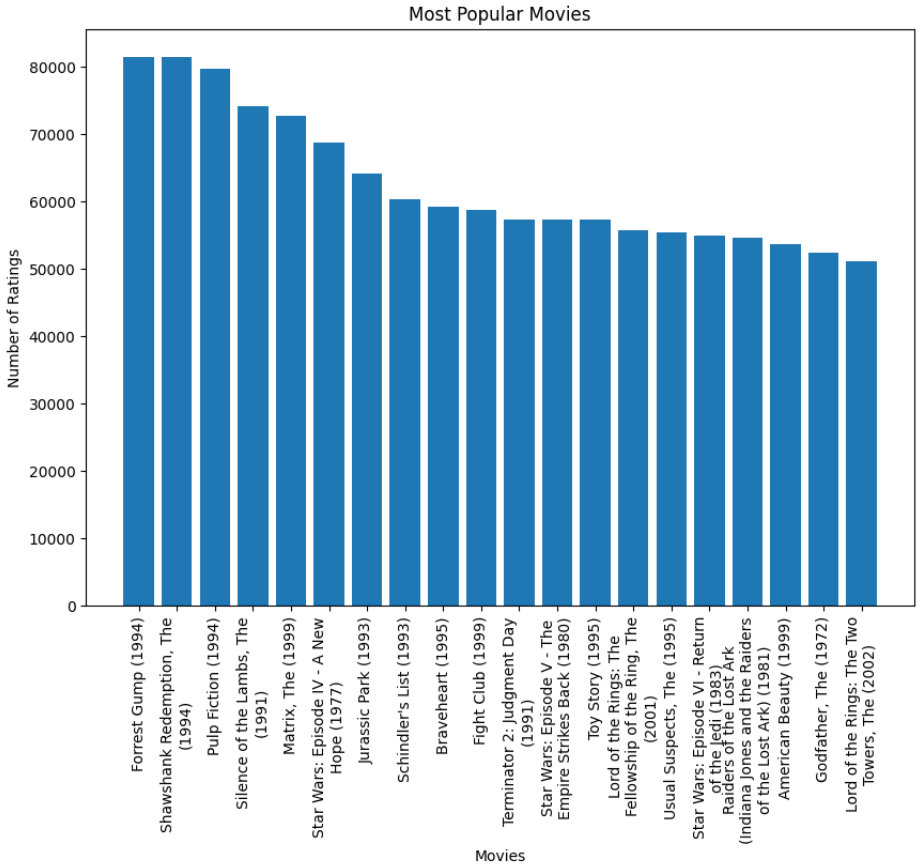

Figure 2: Most Popular Movies vs No. of Ratings

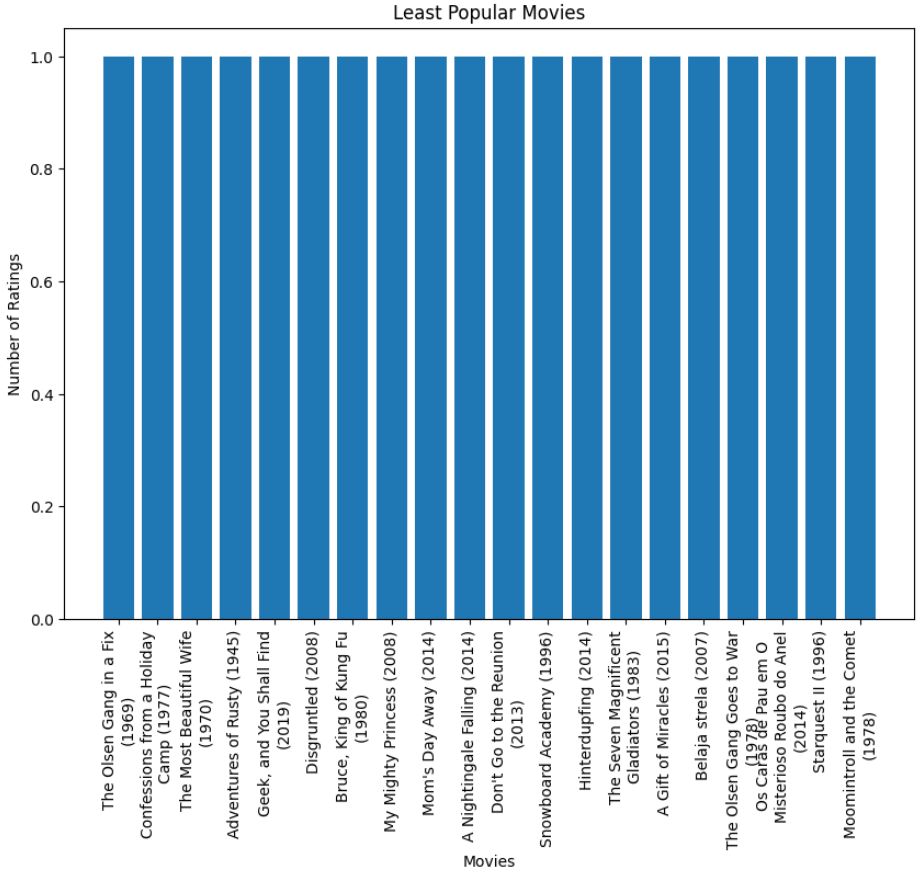

Figure 3: Least Popular Movies vs No. of Ratings

4.2 Exploring Annual Movie Releases

Our research also looked at annual film releases, which provided information on how the business is changing. We aimed to identify patterns and trends by classifying the data annually, which would provide a thorough comprehension of the film industry.

We may also examine the gain percentages by decade to better understand the industry's growth or decline across ten-year periods. Furthermore, examining how well-liked films were during decades may reveal fascinating connections between societal changes and viewer tastes. This thorough methodology offers opportunities for a detailed investigation of societal influences on film consumption in addition to offering a glimpse of cinematic creation across time.

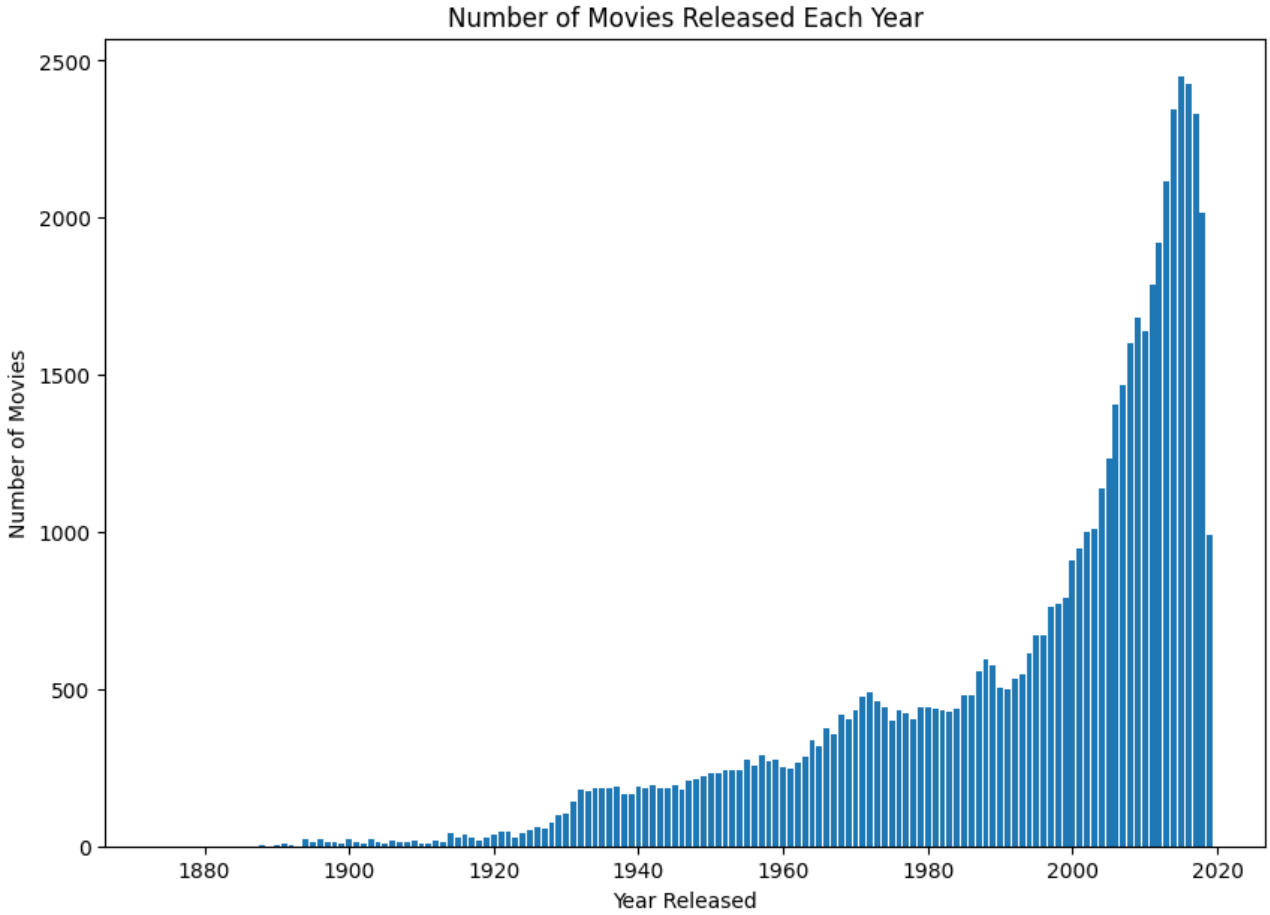

Figure 4: Annual Movie Release

4.3 Analyzing Scatter Plot

The scatter plot depicting the relationship between movie release years and the corresponding number of ratings presents a compelling avenue for exploration. As the trend suggests an increase in ratings with each passing year, several insightful analyses can be pursued.

We can see how audience involvement is changing by examining the ratings trajectory, which also sheds light on how interest has been growing overall over time. Accurately identifying years with very high ratings reveals possible film industry landmarks and helps identify critical junctures and elements that contributed to their success. An in-depth understanding of the complex relationship between public response and praise from critics can be gained by cross-referencing scatter plot data with awards and critical reviews. Furthermore, by using the current dataset for predictive modelling, studios and filmmakers can gain a great deal of insight into future trends in audience interaction.

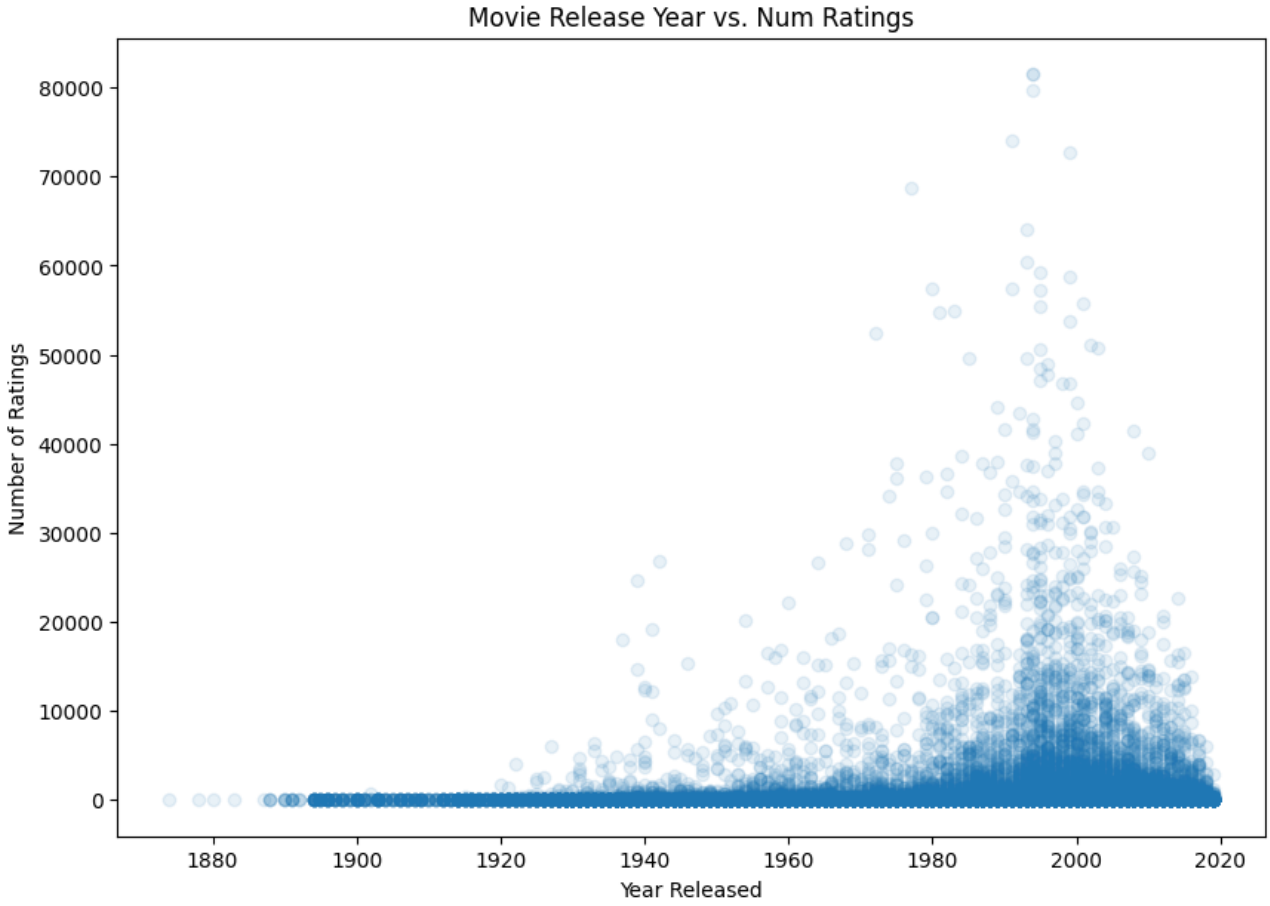

Figure 5: Movie Releases Year vs Number of Ratings

5 Optimizing Movie Recommendations

To initiate our ALS (Alternating Least Squares) model creation, we commence by partitioning our data into training and testing sets with an 80/20% split, setting the seed to 50 for reproducibility. Subsequently, we curate the necessary columns and define a parameter grid using ParamGridBuilder(), encompassing parameters like rank, max iterations, and regularization parameters. This grid facilitates the exploration of various model configurations during the tuning process. Employing the ALS regression model, we navigate through the grid to identify the optimal model fit based on specified parameters such as rank, max iterations, and regularization. Post-identification, we subject the model to rigorous testing to assess its performance. Finally, leveraging the trained model, we generate predictions for movie recommendations, providing valuable insights into user preferences and enhancing the overall recommendation system.

The collaborative filtering method known as ALS matrix factorization is an iterative process aiming to estimate the ratings matrix, denoted as R, by multiplying two lower-rank factor matrices, X and Y (i.e., X * Yt = R). In each iteration, one factor matrix is kept constant, and the other is determined using least squares. This process is alternated until convergence. The model's initialization involves setting the "nonnegative" parameter to "True" to ensure only nonnegative ratings within the 0-5 range are returned. Additionally, the "coldStartStrategy" is set to "drop" to prevent situations where all user's ratings are added to the training set, as this would render predictions meaningless. The "implicitPrefs" parameter is set to "False" since the dataset includes actual movie ratings, making implicit feedback unnecessary.

The ALS method plays a pivotal role in collaborative filtering for recommendation systems, offering diverse applications across various industries. ALS serves as the backbone for crafting personalized recommendation systems, empowering platforms to suggest items based on users' historical preferences, thereby enhancing user experience through tailored content. ALS streamlines the job search process by suggesting suitable job openings to individuals based on their skills, experience, and preferences, thus enhancing the efficiency of job recommendation systems. ALS optimizes targeted advertising by predicting user preferences, allowing for personalized advertisements, and potentially improving engagement and conversion rates.

The ALS method, with its collaborative filtering capabilities, demonstrates versatile and impactful applications across industries, fundamentally reshaping the landscape of recommendation systems. As technology advances, ALS continues to play a vital role in enhancing user experiences and decision-making processes, contributing to the evolution of personalized services in the digital era.

```
# Call the function to get the title for movieId 2906
get_movie_title_from_id(2906)
```

```
'Random Hearts (1999)'
```

Figure 6: Movie Title from Movie ID

```python
import pandas as pd

# Define the file path
file_path = "/content/movies.csv"

# Read CSV file into a DataFrame
movie_titles = spark.read.csv(file_path, header=True, inferSchema=True)

# Show the first few rows of the DataFrame
movie_titles.show(truncate=False)
```

```
+-------+-----------------------------------+-----------------------------------------+
|movieId|title                              |genres                                   |
+-------+-----------------------------------+-----------------------------------------+
|1      |Toy Story (1995)                   |Adventure|Animation|Children|Comedy|Fantasy|
|2      |Jumanji (1995)                     |Adventure|Children|Fantasy                 |
|3      |Grumpier Old Men (1995)            |Comedy|Romance                           |
|4      |Waiting to Exhale (1995)           |Comedy|Drama|Romance                     |
|5      |Father of the Bride Part II (1995) |Comedy                                   |
|6      |Heat (1995)                        |Action|Crime|Thriller                    |
|7      |Sabrina (1995)                     |Comedy|Romance                           |
|8      |Tom and Huck (1995)                |Adventure|Children                       |
|9      |Sudden Death (1995)                |Action                                   |
|10     |GoldenEye (1995)                   |Action|Adventure|Thriller              |
|11     |American President, The (1995)     |Comedy|Drama|Romance                     |
|12     |Dracula: Dead and Loving It (1995) |Comedy|Horror                            |
|13     |Balto (1995)                       |Adventure|Animation|Children           |
|14     |Nixon (1995)                       |Drama                                    |
|15     |Cutthroat Island (1995)            |Action|Adventure|Romance               |
|16     |Casino (1995)                      |Crime|Drama                              |
|17     |Sense and Sensibility (1995)       |Drama|Romance                           |
|18     |Four Rooms (1995)                  |Comedy                                   |
|19     |Ace Ventura: When Nature Calls (1995)|Comedy                                 |
|20     |Money Train (1995)                 |Action|Comedy|Crime|Drama|Thriller     |
+-------+-----------------------------------+-----------------------------------------+
only showing top 20 rows
```

Figure 7: Movie Titles along with Movie ID and Genre

```python
# Call the function to get recommended movies for user 10
get_user_recommended_movies(userRecs, 10)
```

```
Movie:
Zed & Two Noughts, A (1985)
Predicted Rating: 5.8545379638671875

Movie:
Buffalo '66 (a.k.a. Buffalo 66) (1998)
Predicted Rating: 5.813051223754883

Movie:
Emma (2009)
Predicted Rating: 5.685601711273193

Movie:
Pride and Prejudice (1995)
Predicted Rating: 5.434269905090332

Movie:
Superstar (1999)
Predicted Rating: 5.258571147918701
```

Figure 8: User Recommended Movies using User ID

6 Jaccard Index Content-Based Method

The data preprocessing for the Jaccard Index Content-Based method entails a singular primary task: compiling all tags associated with a particular movie into a list. Each movie receives tags from multiple reviewers, and it's plausible that distinct reviewers may assign different tags to the same movie. The objective of this step is to gather all unique tags provided by any user for a movie and concatenate this list of tags sequentially for each movie.

The Jaccard Index Recommendation generated a set of recommended movies by analyzing the tags associated with movies the user previously enjoyed. To enhance the efficacy of this approach, we suggest utilizing more comprehensive datasets containing richer qualitative information. A larger and more high-quality dataset, particularly in terms of tag information, would enable more rigorous Jaccard matching, thereby yielding more reliable and trustworthy recommendations.

7 References

1. **MovieLens 25M Dataset** - https://grouplens.org/datasets/movielens/25m/