gtisnanga /
PHASE-4-PRJOECT-MOVIE-RECOMMENDATIONS

Code Issues Pull requests Actions Projects Wiki Security Insights

0 stars 1 fork 1 watching Branches Tags Activity

Public repository

1 Branch 0 Tags Go to file t Go to file + Add file Code ···

gtisnanga presentations 4158728 · 20 minutes ago

ml-latest-small	Editing README.md	yesterday
Final Notebook.ipynb	Editing README.md	yesterday
Final Notebook.pdf	adding pdf	1 hour ago
Phase 4 Movie Recommenda...	presentations	20 minutes ago
Phase 4 Movie Recommenda...	presentations	20 minutes ago
README.md	Update README.md	2 hours ago

README

Movie Recommendation System for Netflix

Phase: 4 Group: 13

Group Members:

- Sylvia Manono
- Amos Kipngetich
- Angela Maina
- Charles Ndegwa
- Sandra Koech
- Gloria Tisnanga
- Alex Miningwa

Student Pace: Part time

Scheduled Project Review Date/Time: October 14, 2024

Instructor Name: Samuel G. Mwangi

Executive Summary

This project aims to improve Netflix’s recommendation system to solve the issue of users endlessly scrolling to find content they enjoy. By enhancing the recommendation engine, Netflix seeks to provide more personalized and relevant suggestions, increasing user engagement and viewing time, while reducing decision fatigue.

Business and Data Understanding

Netflix, a global streaming platform with a vast and diverse movie catalog, aims to provide personalized content to ensure user engagement and satisfaction. With thousands of options available, users often face difficulty in finding content that matches their preferences, leading to decision fatigue and lower engagement. This results in users endlessly scrolling without finding something to watch quickly.

To address this, Netflix is focused on enhancing its recommendation engine by suggesting films similar to those users have enjoyed, based on the content and genre of the movies. The goal is to build a content-based recommendation system that provides relevant movie suggestions, helping users discover new content easily, stay engaged, and ultimately increase viewing time. This system will enhance Netflix’s ability to deliver a personalized and enjoyable experience for its users.

Dataset Overview

The dataset includes:

Movie Titles, Genres

User IDs, Movie IDs, Ratings This data allows for both content-based and collaborative filtering approaches. EDA Highlights Exploratory data analysis revealed key patterns in user ratings, popular genres, and user-movie interactions. Visualization techniques were used to identify trends and correlations in the data.

Modeling Techniques

Content-Based Filtering: Recommends movies similar to what users have liked using cosine similarity based on genres. Collaborative Filtering: KNNBasic: Predicts ratings based on similar users. KNNWithMeans: Adjusts for user biases by incorporating mean ratings. SVD: Matrix factorization to uncover hidden patterns in user-item interactions. Hybrid Model: Combines content-based and collaborative filtering techniques for enhanced performance.

Objectives

Here are three refined objectives based on your project description:

1. **Develop a Personalized Movie Recommendation System** Build a recommendation model using collaborative filtering (e.g., matrix factorization or deep learning techniques like neural collaborative filtering) to predict user ratings for movies. Recommend the top 5 most relevant movies for each user, enhancing the personalization experience on the platform.

2. **Address Cold Start for New Users** Tackle the cold start issue by implementing content-based filtering and recommending movies based on user preferences (genres, actors). Complement this with trending or popular movies to ensure engagement while collecting more personalized data for future recommendations.

3. **Enhance System Precision and User Feedback Integration** Improve recommendation accuracy and relevance through a hybrid approach that combines collaborative filtering and content-based filtering. Implement a feedback mechanism to refine the model continuously, using metrics such as RMSE, Precision@K, and F1 Score for evaluation, and leverage user ratings to enhance future recommendations.

DATA DESCRIPTION

There are a number of csv files available with different columns in the Data file.

movies.csv

movieId - Unique identifier for each movie.

title - The movie titles.

genre - The various genres a movie falls into.

ratings.csv

userId - Unique identifier for each user

movieId - Unique identifier for each movie.

rating - A value between 0 to 5 that a user rates a movie on. 5 is the highest while 0 is the lowest rating.

timestamp - This are the seconds that have passed since Midnight January 1, 1970(UTC)

tags.csv

userId - Unique identifier for each user

movieId - Unique identifier for each movie.

tag - A phrase determined by the user.

timestamp - This are the seconds that have passed since Midnight January 1, 1970(UTC)

links.csv

movieId - It's an identifier for movies used by https://movielens.org and has link to each movie.

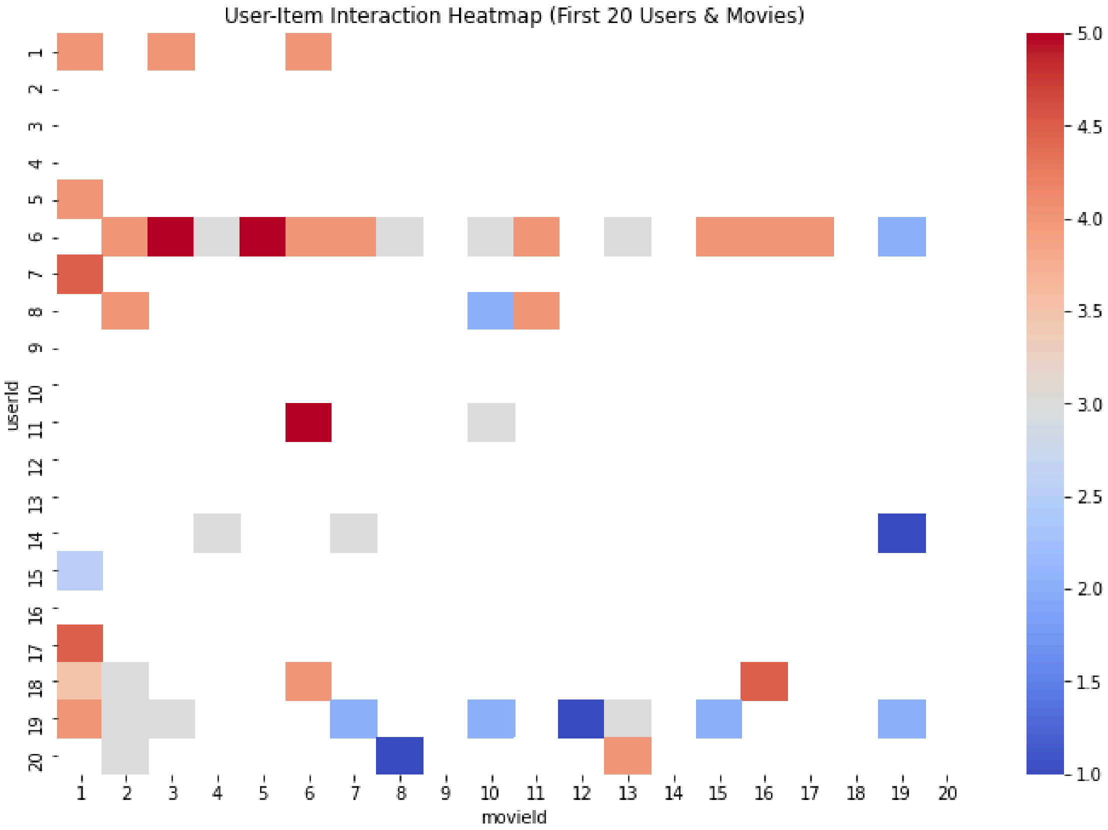
imdbId - It's an identifier for movies used by http://www.imdb.com and has link to each movie.

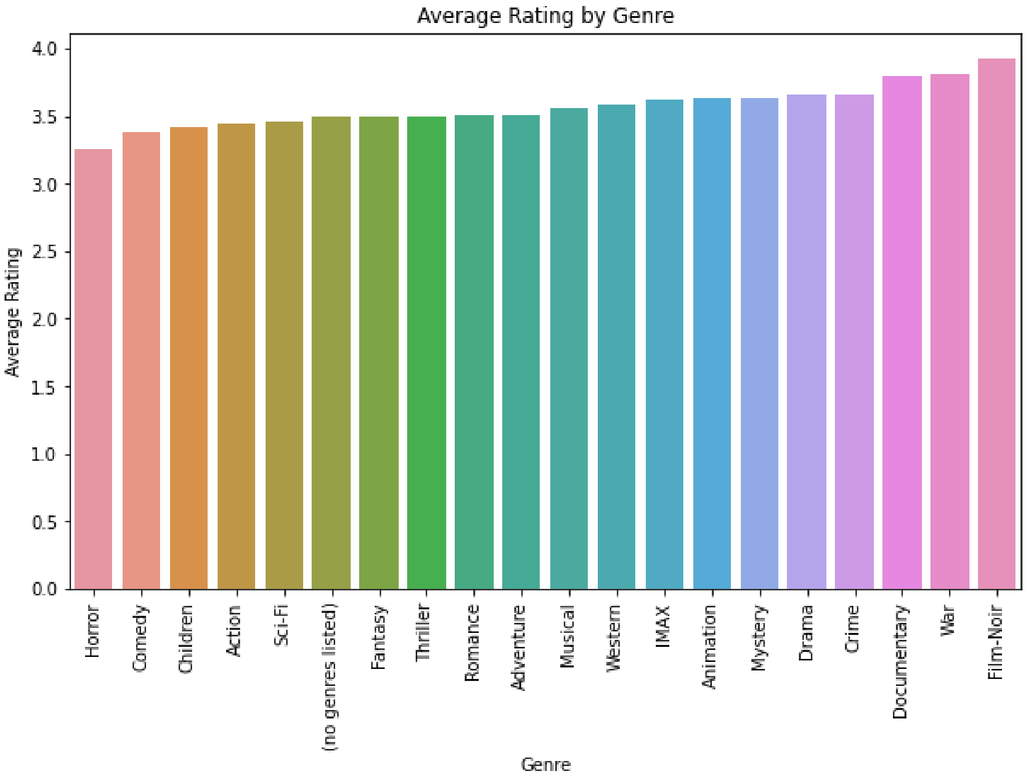
tmdbId - is an identifier for movies used by https://www.themoviedb.org and has link to each movie.

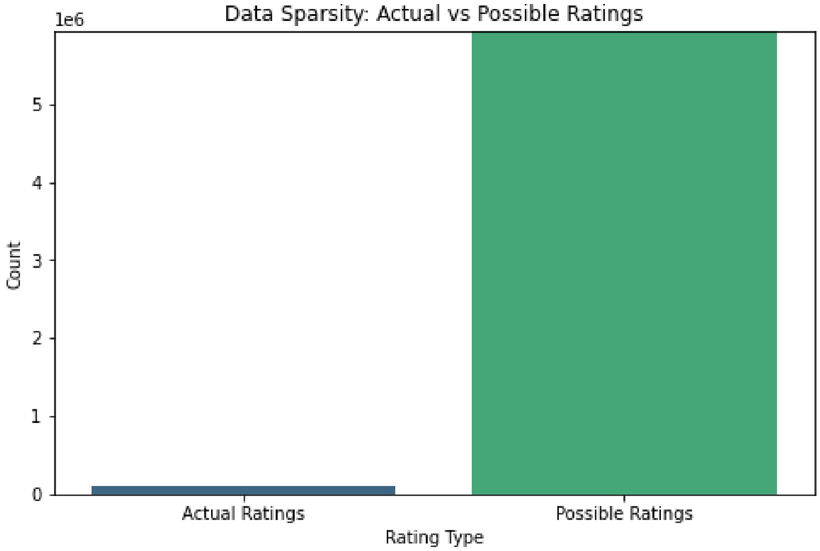
EDA

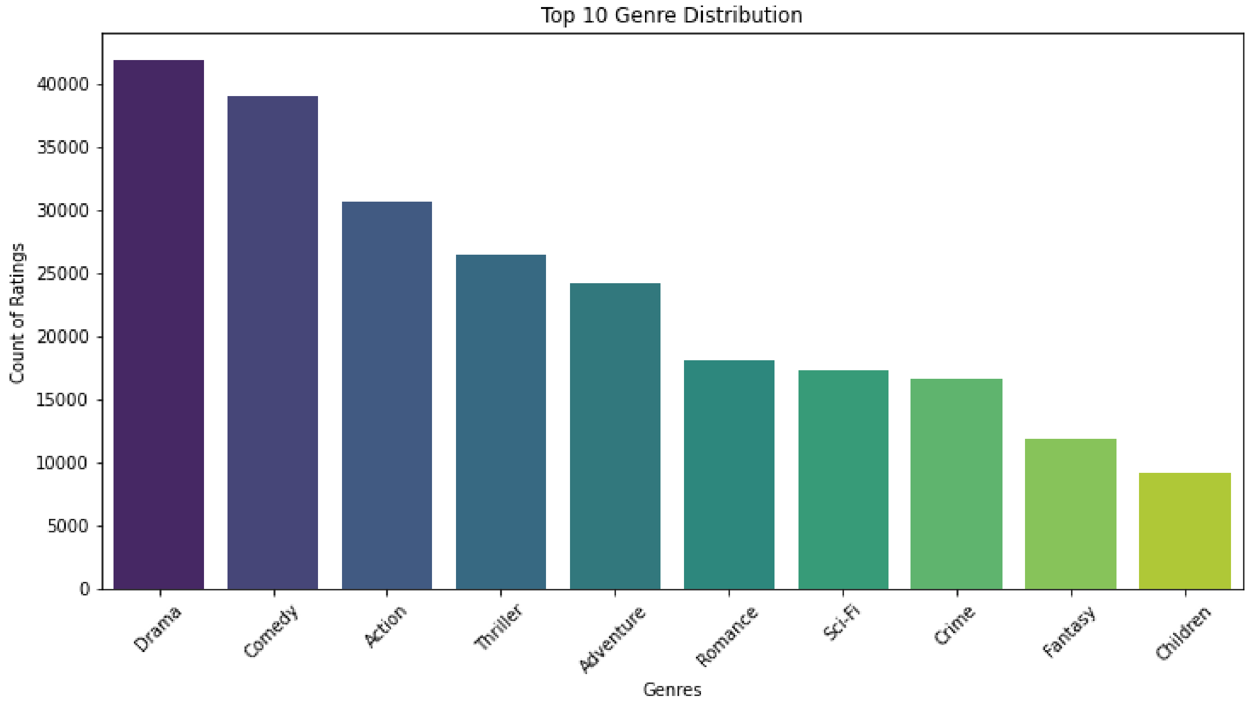
Leading Questions

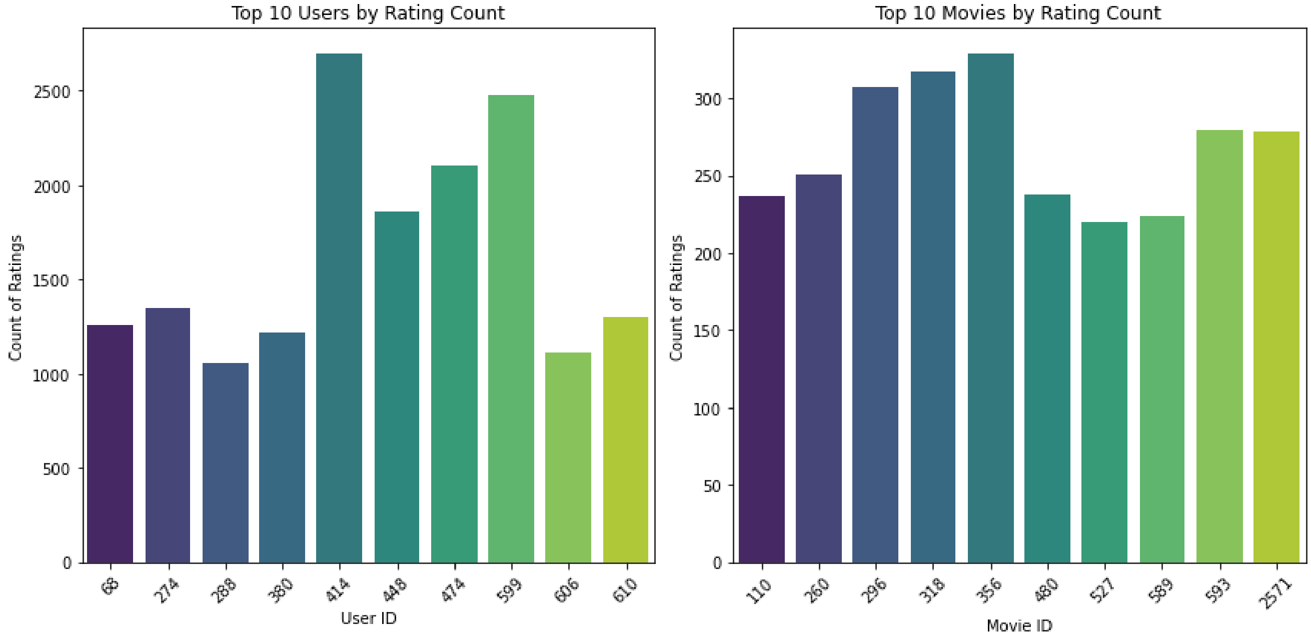
1. User Activity Levels: Which users are the most active, and how does their activity compare to less active users?
2. Item Popularity: Which movies are rated most frequently, and are there patterns in terms of genre or release year?
3. Rating Trends Over Time: How do average ratings and the number of ratings submitted change over time, and are there any seasonal or genre-specific trends?



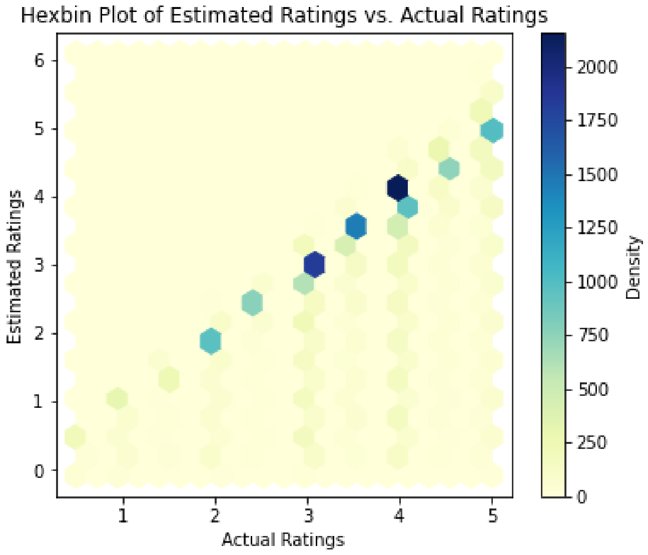

Average Rating by Genre
Average Rating
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Horror
Comedy
Children
Action
Sci-Fi
(no genres listed)
Fantasy
Thriller
Romance
Adventure
Musical
Western
IMAX
Animation
Mystery
Drama
Crime
Documentary
War
Film-Noir
Genre


Data Sparsity: Actual vs Possible Ratings
1e6
Count
5
4
3
2
1
0
Actual Ratings
Possible Ratings
Rating Type


Top 10 Genre Distribution
Count of Ratings
0
5000
10000
15000
20000
25000
30000
35000
40000
Drama
Comedy
Action
Thriller
Adventure
Romance
Sci-Fi
Crime
Fantasy
Children
Genres


Top 10 Users by Rating Count
Count of Ratings
0
500
1000
1500
2000
2500
68
274
288
380
414
448
474
599
606
610
User ID
Top 10 Movies by Rating Count
Count of Ratings
0
50
100
150
200
250
300
110
260
296
318
356
480
527
589
593
2571
Movie ID

Modelling



Value of 1.1689: This specific RMSE value means that, on average, the estimated ratings from your deployed model deviate from the actual ratings by approximately 1.17 units on the rating scale. Scale Context: If your ratings are on a scale of 1 to 5, an RMSE of about 1.17 suggests that the predictions are relatively close to the actual ratings. This is a better performance compared to the RMSE of 2.5133 from the hybrid model you evaluated earlier. It implies that the deployed model provides more accurate predictions than the hybrid model.

Analyzing the Updated Hexbin Plot Key Observations:

Strong Positive Correlation: The plot clearly shows a strong positive correlation between actual and estimated ratings, as evidenced by the diagonal line of dense hexagons. This indicates that the model's predictions generally align with the actual ratings.

Clustering: The data points cluster tightly along the diagonal line, suggesting that the model's predictions are consistently accurate.

Density: The color intensity of the hexagons reveals the density of data points. The darker blue hexagons indicate areas with a higher concentration of data points, suggesting that the model's predictions are more frequent in those ranges of actual and estimated ratings.

Sparse Areas: There are some sparse areas, especially in the lower-left and upper-right corners, indicating fewer data points in those regions. This might be due to a lack of training data or inherent limitations in the model's ability to predict ratings in those areas.

Conclusion

The evaluation of your movie recommendation system reveals promising results. The RMSE of **1.1689** indicates that the deployed model's predictions are closely aligned with actual ratings, enhancing the overall accuracy compared to the previous hybrid model's RMSE of **2.5133**. This suggests that the refined approach effectively captures user preferences, contributing to a more personalized experience. The hexbin plot analysis further demonstrates a strong positive correlation between actual and estimated ratings, highlighting the model's reliability and consistency in predicting user ratings.

Recommendations

1. **Refine the Collaborative Filtering Approach**:
   - Continue enhancing the collaborative filtering model by experimenting with different matrix factorization techniques and tuning hyperparameters to further reduce RMSE and improve accuracy.
2. **Leverage Content-Based Filtering**:
   - Implement content-based filtering techniques for new users to alleviate the cold start problem. Use

Releases

No releases published
Create a new release

Packages

No packages published
Publish your first package

Contributors 4

gtisnanga

ANGI3MAINA Angela W Maina

Rooszta

Roticha

Languages

● Jupyter Notebook 100.0%