Movie Recommendation System

DSC 680 Applied Data Science

Bellevue University

Spring 2021

Conrad Ibañez

**Abstract**

When it comes to entertainment and the programs we watch, we are like kids in a candy store. With the growth of online streaming, there are so many choices for movies and shows that are available at our fingertips. Recommendation systems have become a common feature in many e-commerce sites including streaming services such as Netflix or YouTube. Users interact with these systems to have more personalized experiences and to explore new content. Recommendation systems are based on different algorithms and filtering methods, such as collaborative filtering and content-based filtering. They factor in different sources of information to provide recommendations for users. This project implements a movie recommendation system.

**Background**

In today’s digital world, consumers have many choices. Whether you are browsing the web, shopping for items, or watching online content, you have likely interacted with a recommendation system. Recommendation systems suggest items that may be of interest to you. For example, when you are shopping for a smartphone, an e-commerce site may suggest compatible accessories such as a case or headset that you may want to include in your purchase. If you have watched content on Netflix or YouTube, both applications have items in queue to watch next.

The designs for recommendation systems mimic how people would choose candy in a candy store. Individuals would rely on current knowledge, preferences, or past experiences, such as which candy they tried and liked or did not like. Additionally, people make decisions based on input from others around them such as family and friends.

There are many approaches in the design of recommendation systems. One common approach is collaborative filtering. Collaborative filtering makes recommendations based on patterns of ratings or usage. Another approach is content-based filtering where a system determines what a user would like by analyzing information describing the item while factoring in the user’s preferences. Other approaches for recommendations systems combine different aspects to form a hybrid approach.

**Problem Statement**

Consumers of online content are constantly looking for entertainment options. It is important that movie recommendation systems work well so that customers are satisfied with the movies that they watch. Many of us have probably wasted time cycling through movies and watching the beginning parts only to stop and choose another movie. A movie recommendation system that works well leads to many satisfied customers who feel that there are endless movies to watch. They become loyal to a business and are dependent on the services that are provided.

This project will analyze movies and reviews data and propose a movie recommendation system. It will investigate the different approaches that are used by recommendation systems. This project will address various questions.

1. What data was available and how was it used in the recommendation system?
2. What limitations or problems existed because of the data?
3. How do the different approaches of recommendation systems work?
4. What are the benefits versus the disadvantages the approaches?
5. Does one approach perform better than the others?
6. Is there a better approach that can be created?
7. How does your recommendation system work?
8. What was the output/recommendations given?
9. What issues did you face?
10. What improvements can you make to your implementation?

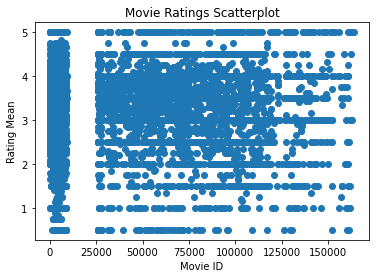
**Data Understanding**

The dataset for this project is The Movies Dataset found on Kaggle- https://www.kaggle.com/rounakbanik/the-movies-dataset. This dataset contains metadata for the Full MovieLens Dataset (https://grouplens.org/datasets/movielens/latest/), which contains over 45,000 movies and over 26 million ratings from more than 270,000 users. The Movies Dataset contains seven csv files of data out of which two files are subset data files.

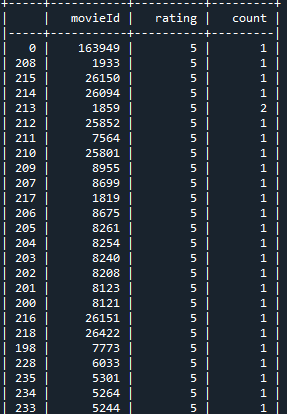
**Methods**

Exploratory data analysis was done to understand the data that was available for the recommendation engine. There were only minor issues with missing or null data in the data set for the attributes that were used. Some movies had missing data and therefore were not factored in for the recommendations. Relevant data for movies were found in multiple tables. There were different id’s for each of the movies, so determining which key to join the tables was important.

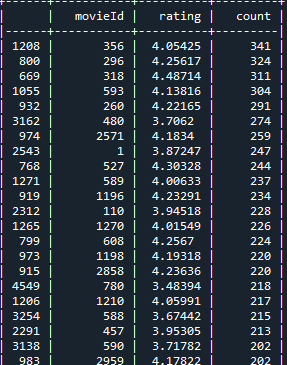
One approach for the recommendation system is to recommend the top movies with the highest rating. One problem with this approach is that the dataset has over 45,000 movies, and even the smaller dataset is still relatively large with over 9k movies. The scatterplot below of movies with their rating mean is not very valuable because of the large data.



To narrow the amount of data, we look to set a minimum threshold for the number of reviews. This will prevent movies with only a few reviews and high rating mean from being the first to be recommended as we can see many movies with the highest rating (5) with only 1 review.



We can see other movies with many reviews and relatively high ratings but not the topmost rating.



Another approach we want to investigate in this project is to evaluate the patterns of the reviews of certain users. If we can detect a pattern where a user is giving high reviews for certain sets of movies, then we can factor this method in improving the recommendation system. This work is still pending.

For this project, the full set of ratings by different users were evaluated to form a weighted rating value for the recommendation system. This is different from the other projects that were proposed by others who have used this data. For the most part, they used the vote\_average and vote\_count in the metadata for the movies to form their rating system. Additionally for this project, genres were evaluated extensively. Movies were classified into different genres or aggregate genres, where movies could be long in many genres. Again this is different from other researchers in that they primarily used one genre for classification of each move.

**Results and Conclusion**

In this project, the 75th quantile was used to get a subset of data composed of 11,414 movies to be considered in the recommendations based on the number of ratings for the movies and also using a weighted rating value. The formula for the weighted rating is the following:

(v/(v+m) \* R) + (m/(m+v) \* C)

where,

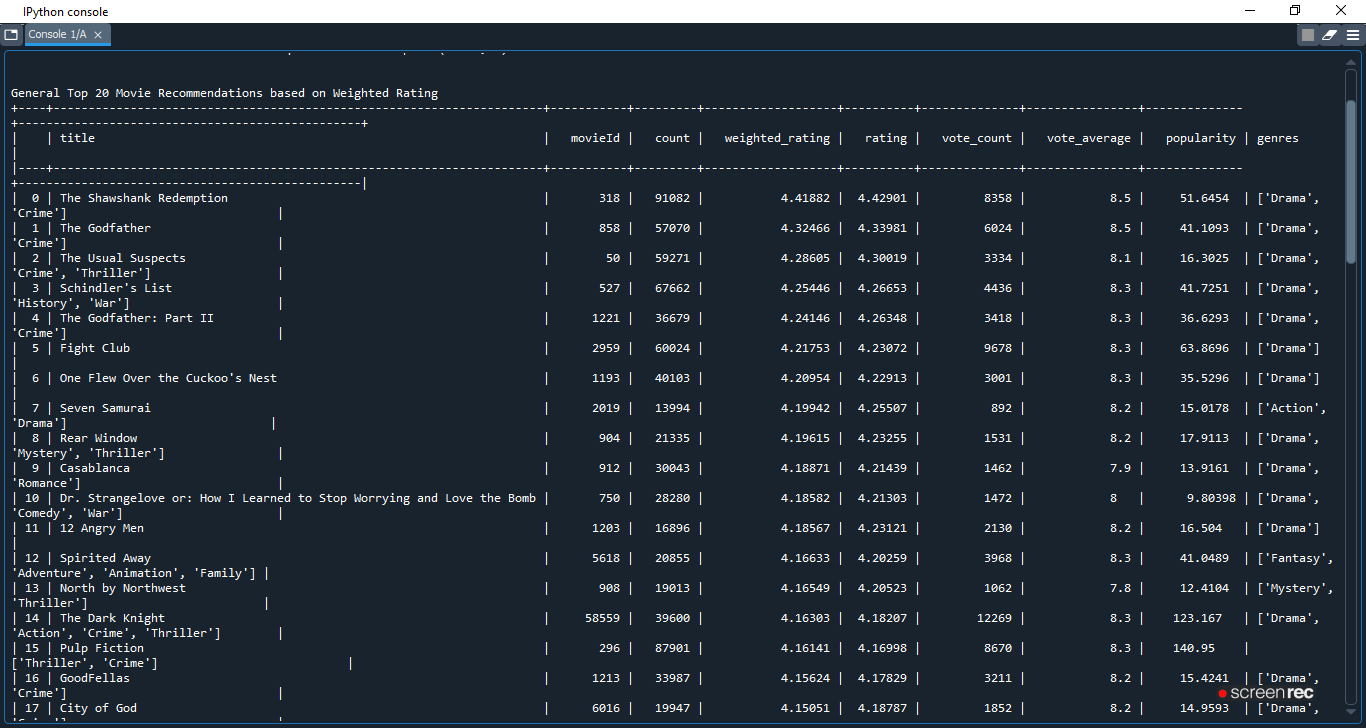
v is the number of ratings for the movie

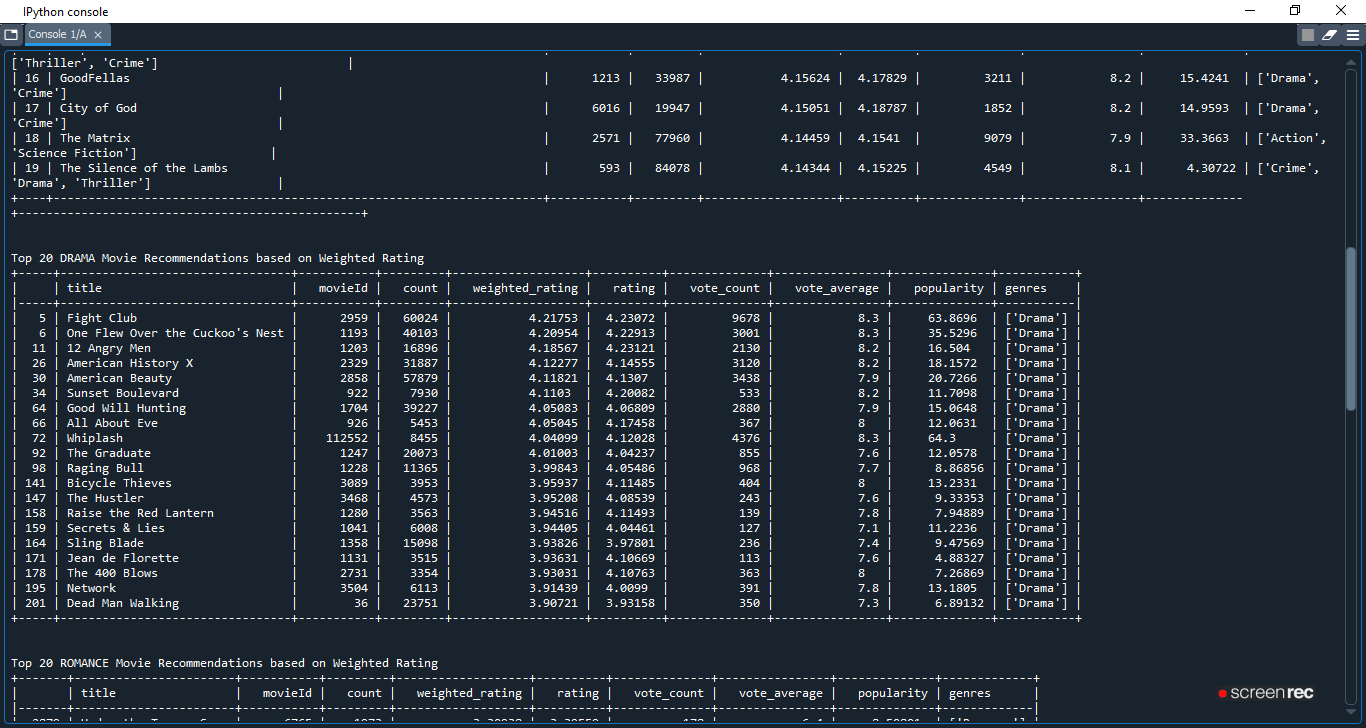
m is the minimum ratings required to be listed in the chart

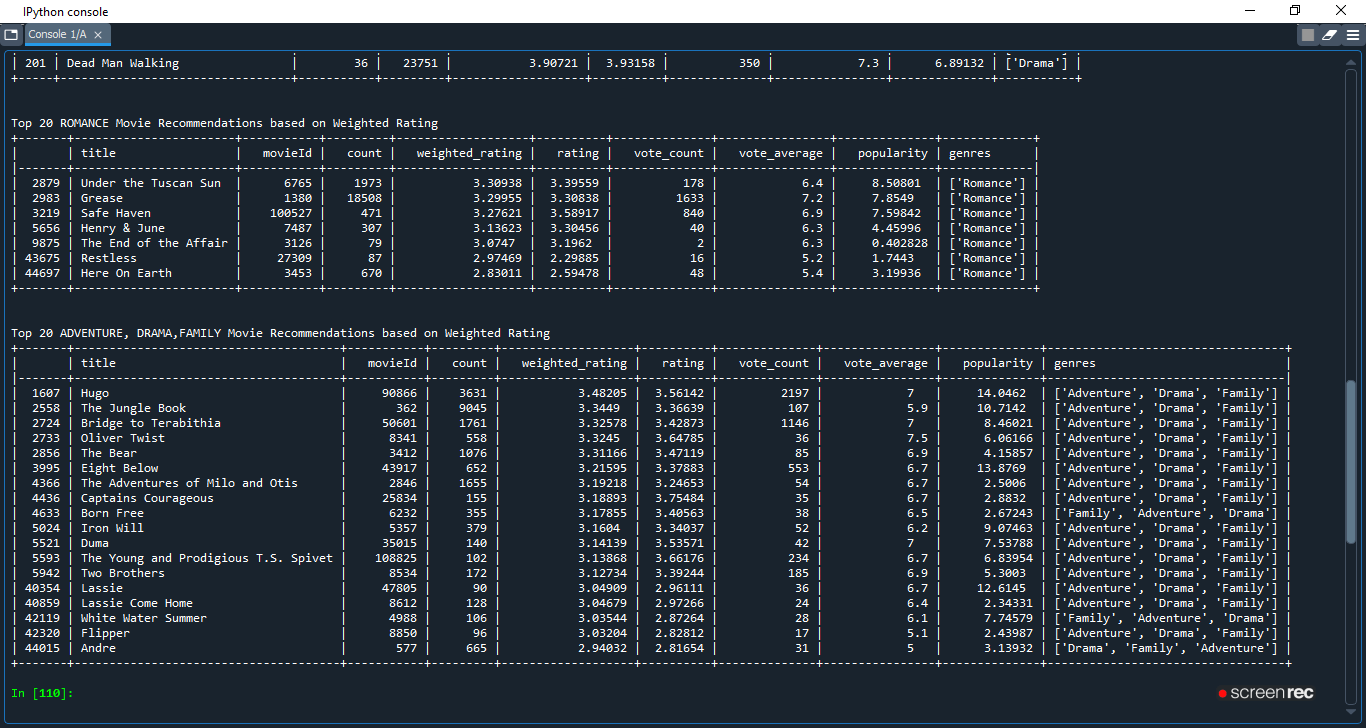
R is the average rating of the movie

C is the mean rating for the data set

The following recommendations were generated by the system:







During testing of the output, there were times where not many recommendations showed up for the different genres. Therefore the quantile was reduced from 90th gradually to 75th. For less popular genres, the quantile may need to be further adjusted to have some recommendations appear. In order to test the effectiveness of this recommendation system, it should be deployed and customer actions should be monitored to determine if they watch movies based on the recommendations.

**Acknowledgements**

Thank you to the class and professor for guidance and input for this project. Also, thank you to the researchers listed in the references for their ground work. Some of their code were used or incorporated into this project.

**References**

1. <https://www.kaggle.com/rounakbanik/the-movies-dataset> – Kaggle dataset which is the primary dataset for this project.
2. <https://www.kaggle.com/rounakbanik/movie-recommender-systems> - Recommender system project by the same user for the dataset
3. <https://www.kaggle.com/ibtesama/getting-started-with-a-movie-recommendation-system> - Another project for the dataset
4. <https://grouplens.org/datasets/movielens/latest/> - The full dataset on which the Kaggle dataset is based.
5. Bobadilla J., Ortega F., Hernando A., Gutierrez A., “Recommender systems survey,” Knowledge-Based Systems 46 109-132, 2013. – Discusses different research and approaches to recommendation systems.
6. Bennett J., Lanning S., "The Netflix Prize," Proceedings of KDD Cup and Workshop 2007, 2007.- Competition that sparked the interest for this project.
7. Covington P., Adams J., Sargin E., "Deep Neural Networks for YouTube Recommendations," Proceedings of the 10th ACM Conference on Recommender Systems, 2016. – Research paper for neural network
8. Harper F.M., Konstan J., "The MovieLens Datasets: History and Context," ACM Trans. Interact. Intell. Syst. V, N, Article XXXX, 2015. – More details on the MovieLens dataset
9. Koren Y., Bell R., "Advances in Collaborative Filtering," Recommender Systems Handbook, 2011. – Discusses collaborative filtering
10. Smith B., Linden G., “Two Decades of Recommender Systems at Amazon.com,” IEEE Internet Computing, 2017. – Discusses Amazon deployment of their recommendation systems.