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?

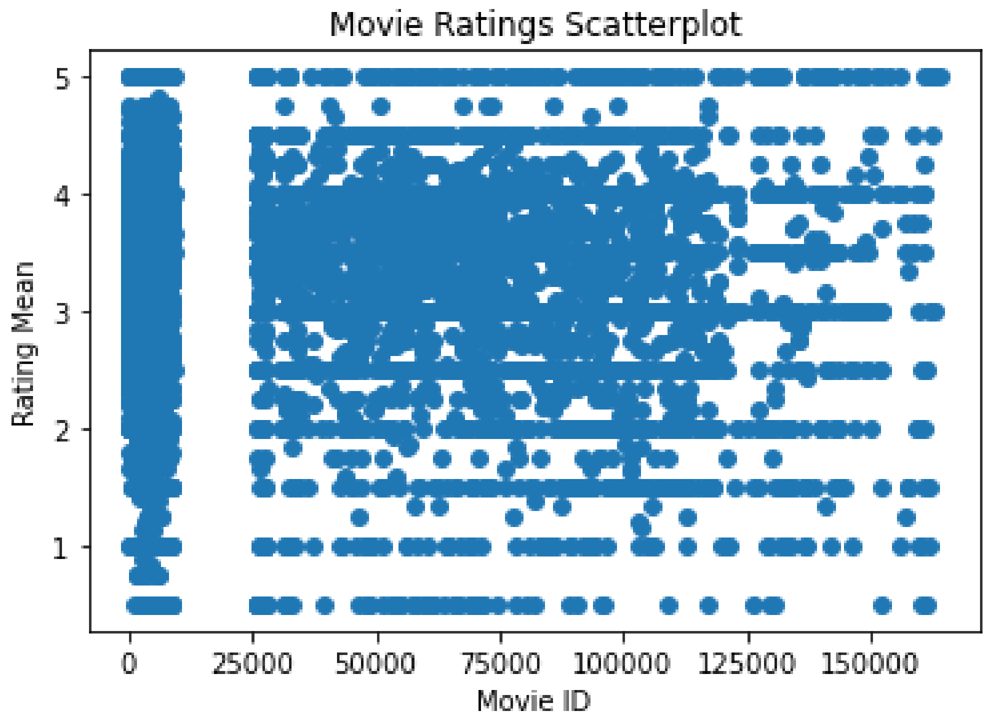
Data Understanding

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.

	movieId	rating	count
0	163949	5	1
208	1933	5	1
215	26150	5	1
214	26094	5	1
213	1859	5	2
212	25852	5	1
211	7564	5	1
210	25801	5	1
209	8955	5	1
207	8699	5	1
217	1819	5	1
206	8675	5	1
205	8261	5	1
204	8254	5	1
203	8240	5	1
202	8208	5	1
201	8123	5	1
200	8121	5	1
216	26151	5	1
218	26422	5	1
198	7773	5	1
228	6033	5	1
235	5301	5	1
234	5264	5	1
233	5244	5	1

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

	movieId	rating	count
1208	356	4.05425	341
800	296	4.25617	324
669	318	4.48714	311
1055	593	4.13816	304
932	260	4.22165	291
3162	480	3.7062	274
974	2571	4.1834	259
2543	1	3.87247	247
768	527	4.30328	244
1271	589	4.00633	237
919	1196	4.23291	234
2312	110	3.94518	228
1265	1270	4.01549	226
799	608	4.2567	224
973	1198	4.19318	220
915	2858	4.23636	220
4549	780	3.48394	218
1206	1210	4.05991	217
3254	588	3.67442	215
2291	457	3.95305	213
3138	590	3.71782	202
983	2959	4.17822	202

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:

General Top 20 Movie Recommendations based on Weighted Rating

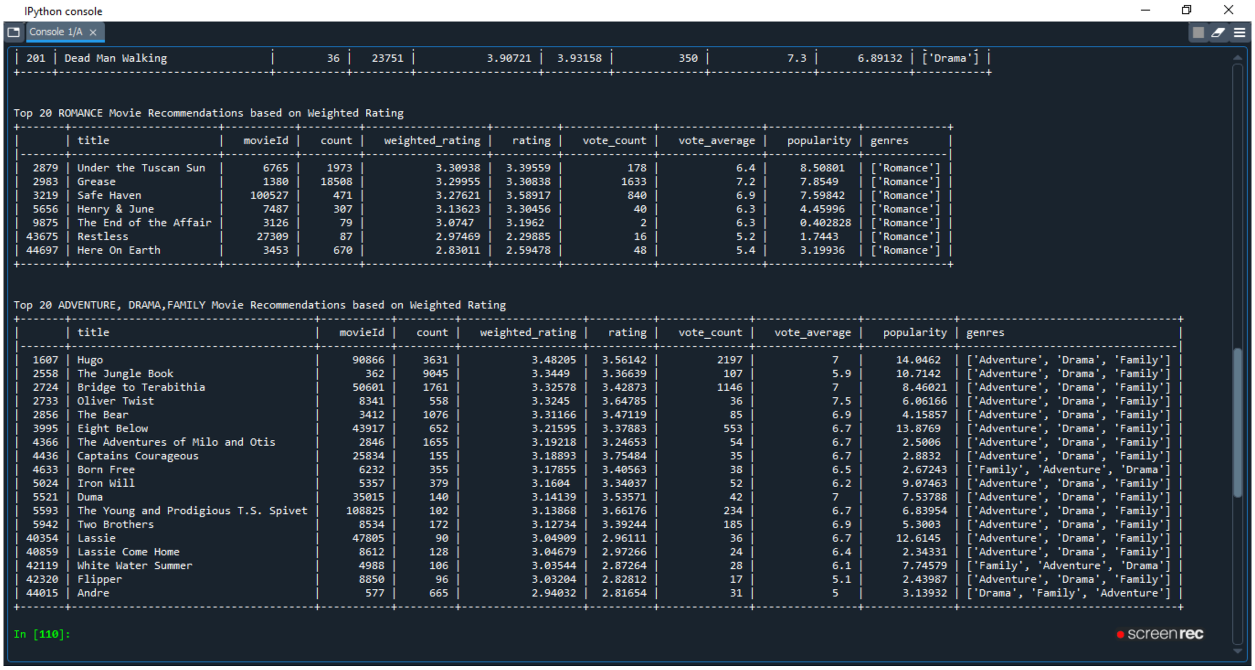
	title	movieId	count	weighted_rating	rating	vote_count	vote_average	popularity	genres
0	The Shawshank Redemption	318	91082	4.41882	4.42901	8358	8.5	51.6454	['Drama', 'Crime']
1	The Godfather	858	57070	4.32466	4.33981	6024	8.5	41.1093	['Drama', 'Crime']
2	The Usual Suspects	50	59271	4.28605	4.30019	3334	8.1	16.3025	['Drama', 'Crime', 'Thriller']
3	Schindler's List	527	67662	4.25446	4.26653	4436	8.3	41.7251	['Drama', 'History', 'War']
4	The Godfather: Part II	1221	36679	4.24146	4.26348	3418	8.3	36.6293	['Drama', 'Crime']
5	Fight Club	2959	60024	4.21753	4.23072	9678	8.3	63.8696	['Drama']
6	One Flew Over the Cuckoo's Nest	1193	40103	4.20954	4.22913	3001	8.3	35.5296	['Drama']
7	Seven Samurai	2019	13994	4.19942	4.25507	892	8.2	15.0178	['Action', 'Drama']
8	Rear Window	904	21335	4.19615	4.23255	1531	8.2	17.9113	['Drama', 'Mystery', 'Thriller']
9	Casablanca	912	30043	4.18871	4.21439	1462	7.9	13.9161	['Drama', 'Romance']
10	Dr. Strangelove or: How I Learned to Stop Worrying and Love the Bomb	750	28280	4.18582	4.21303	1472	8	9.80398	['Drama', 'Comedy', 'War']
11	12 Angry Men	1203	16896	4.18567	4.23121	2130	8.2	16.504	['Drama']
12	Spirited Away	5618	20855	4.16633	4.20259	3968	8.3	41.0489	['Fantasy', 'Adventure', 'Animation', 'Family']
13	North by Northwest	908	19013	4.16549	4.20523	1062	7.8	12.4104	['Mystery', 'Thriller']
14	The Dark Knight	58559	39600	4.16303	4.18207	12269	8.3	123.167	['Drama', 'Action', 'Crime', 'Thriller']
15	Pulp Fiction	296	87901	4.16141	4.16998	8670	8.3	140.95	['Thriller', 'Crime']
16	GoodFellas	1213	33987	4.15624	4.17829	3211	8.2	15.4241	['Drama', 'Crime']
17	City of God	6016	19947	4.15051	4.18787	1852	8.2	14.9593	['Drama', 'Crime']
18	The Matrix	2571	77960	4.14459	4.1541	9079	7.9	33.3663	['Action', 'Science Fiction']
19	The Silence of the Lambs	593	84078	4.14344	4.15225	4549	8.1	4.30722	['Crime', 'Drama', 'Thriller']

Top 20 DRAMA Movie Recommendations based on Weighted Rating

	title	movieId	count	weighted_rating	rating	vote_count	vote_average	popularity	genres
5	Fight Club	2959	60024	4.21753	4.23072	9678	8.3	63.8696	['Drama']
6	One Flew Over the Cuckoo's Nest	1193	40103	4.20954	4.22913	3001	8.3	35.5296	['Drama']
11	12 Angry Men	1203	16896	4.18567	4.23121	2130	8.2	16.504	['Drama']
26	American History X	2329	31887	4.12277	4.14555	3120	8.2	18.1572	['Drama']
30	American Beauty	2858	57879	4.11821	4.1307	3438	7.9	20.7266	['Drama']
34	Sunset Boulevard	922	7930	4.1103	4.20082	533	8.2	11.7098	['Drama']
64	Good Will Hunting	1704	39227	4.05083	4.06809	2880	7.9	15.0648	['Drama']
66	All About Eve	926	5453	4.05045	4.17458	367	8	12.0631	['Drama']
72	Whiplash	112552	8455	4.04099	4.12028	4376	8.3	64.3	['Drama']
92	The Graduate	1247	20073	4.01003	4.04237	855	7.6	12.0578	['Drama']
98	Raging Bull	1228	11365	3.99843	4.05486	968	7.7	8.86856	['Drama']
141	Bicycle Thieves	3089	3953	3.95937	4.11485	404	8	13.2331	['Drama']
147	The Hustler	3468	4573	3.95208	4.08539	243	7.6	9.33353	['Drama']
158	Raise the Red Lantern	1280	3563	3.94516	4.11493	139	7.8	7.94889	['Drama']
159	Secrets & Lies	1041	6008	3.94405	4.04461	127	7.1	11.2236	['Drama']
164	Sling Blade	1358	15098	3.93826	3.97801	236	7.4	9.47569	['Drama']
171	Jean de Florette	1131	3515	3.93631	4.10669	113	7.6	4.88327	['Drama']
178	The 400 Blows	2731	3354	3.93031	4.10763	363	8	7.26869	['Drama']
195	Network	3504	6113	3.91439	4.0099	391	7.8	13.1805	['Drama']
201	Dead Man Walking	36	23751	3.90721	3.93158	350	7.3	6.89132	['Drama']

Top 20 ROMANCE Movie Recommendations based on Weighted Rating

	title	movieId	count	weighted_rating	rating	vote_count	vote_average	popularity	genres



IPython console

Console 1/A

201	Dead Man Walking	36	23751	3.90721	3.93158	350	7.3	6.89132	['Drama']

Top 20 ROMANCE Movie Recommendations based on Weighted Rating

	title	movieId	count	weighted_rating	rating	vote_count	vote_average	popularity	genres
2879	Under the Tuscan Sun	6765	1973	3.30938	3.39559	178	6.4	8.50801	['Romance']
2983	Grease	1380	18508	3.29955	3.30838	1633	7.2	7.8549	['Romance']
3219	Safe Haven	100527	471	3.27621	3.58917	840	6.9	7.59842	['Romance']
5656	Henry & June	7487	307	3.13623	3.30456	40	6.3	4.45996	['Romance']
9875	The End of the Affair	3126	79	3.0747	3.1962	2	6.3	0.402828	['Romance']
43675	Restless	27309	87	2.97469	2.29885	16	5.2	1.7443	['Romance']
44697	Here On Earth	3453	670	2.83011	2.59478	48	5.4	3.19936	['Romance']

Top 20 ADVENTURE, DRAMA,FAMILY Movie Recommendations based on Weighted Rating

	title	movieId	count	weighted_rating	rating	vote_count	vote_average	popularity	genres
1607	Hugo	90866	3631	3.48205	3.56142	2197	7	14.0462	['Adventure', 'Drama', 'Family']
2558	The Jungle Book	362	9045	3.3449	3.36639	107	5.9	10.7142	['Adventure', 'Drama', 'Family']
2724	Bridge to Terabithia	50601	1761	3.32578	3.42873	1146	7	8.46021	['Adventure', 'Drama', 'Family']
2733	Oliver Twist	8341	558	3.3245	3.64785	36	7.5	6.06166	['Adventure', 'Drama', 'Family']
2856	The Bear	3412	1076	3.31166	3.47119	85	6.9	4.15857	['Adventure', 'Drama', 'Family']
3995	Eight Below	43917	652	3.21595	3.37883	553	6.7	13.8769	['Adventure', 'Drama', 'Family']
4366	The Adventures of Milo and Otis	2846	1655	3.19218	3.24653	54	6.7	2.5006	['Adventure', 'Drama', 'Family']
4436	Captains Courageous	25834	155	3.18893	3.75484	35	6.7	2.8832	['Adventure', 'Drama', 'Family']
4633	Born Free	6232	355	3.17855	3.40563	38	6.5	2.67243	['Family', 'Adventure', 'Drama']
5024	Iron Will	5357	379	3.1604	3.34037	52	6.2	9.07463	['Adventure', 'Drama', 'Family']
5521	Duma	35015	140	3.14139	3.53571	42	7	7.53788	['Adventure', 'Drama', 'Family']
5593	The Young and Prodigious T.S. Spivet	108825	102	3.13868	3.66176	234	6.7	6.83954	['Adventure', 'Drama', 'Family']
5942	Two Brothers	8534	172	3.12734	3.39244	185	6.9	5.3003	['Adventure', 'Drama', 'Family']
40354	Lassie	47805	90	3.04909	2.96111	36	6.7	12.6145	['Adventure', 'Drama', 'Family']
40859	Lassie Come Home	8612	128	3.04679	2.97266	24	6.4	2.34331	['Adventure', 'Drama', 'Family']
42119	White Water Summer	4988	106	3.03544	2.87264	28	6.1	7.74579	['Family', 'Adventure', 'Drama']
42320	Flipper	8850	96	3.03204	2.82812	17	5.1	2.43987	['Adventure', 'Drama', 'Family']
44015	Andre	577	665	2.94032	2.81654	31	5	3.13932	['Drama', 'Family', 'Adventure']

In [110]:

screenrec

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.