Which Movie to Watch?

Utilizing Collaborative Filtering to Make the Recommendation System

Lifu “Mario” Ma

Yifan Zhu

Xinyu Liu

Sungho Lee

Table of Contents

1. Summary

How to make recommendations to customers is always one of the top questions faced by many companies. For some media-related companies like Netflix, Youtube, Tik-Tok, and some ecommerce companies like Amazon, they have devoted enormous amount of money to develop a recommendation system to attract more customers and keep customer loyalty.

In this report, we take the movie-recommendation problem as an example, use different ways of collaborative filtering to predict and recommend three movies to our group members based on rating information collected from our classmates. Moreover, we try to solve the problem of predicting ratings for new movies without historical ratings by adding more movie-related information like movie type, release of year, movie length, and adding some classmates' information like gender to make predictions. Furthermore, we help make non-personalized recommendations to three brand-new customers and compare our prediction results with more movie rating information that we get from these new customers.

However, we still have a lot of deficiencies. We cannot measure our accuracy of prediction since we don't have our actual ratings. What's more, the non-personalized recommendations for new customers are still not persuasive enough. To solve these problems, we plan to compare our ratings after we watch these movies to check the accuracy, and collect more background information like age group, ethnicity, culture background, personal preferences of new customers to make more precise recommendations.

2. Introduction

It is always a problem for people when deciding which movie to watch, because there are so many choices in Netflix or Youtube. Also, the quality of the movie may also be one problem since no one wants to waste one or two hours and get nothing. To solve these problems, in this report, we use collaborative filtering to predict our group members' ratings and recommend three 3 movies of our choice. We also figure out ways to predict the ratings for new movies and new people, and discuss the value of our recommendation engine.

3. Problem Formulation

To better recommend movies to our group members and predict our ratings, we decide to first select three movies based on our interest, and then calculate the degree of similarity between group members and all the classmates. After that, we could calculate our predictive ratings based on similarity.

For new movies haven't rated by classmates, first we can find people who have watched these movies in DBMI data and calculate our ratings based on similarity between these people and us. We also collect information about the category, year of release and length of movies to make content-based predictions.

For new people without any prior information, we decide to apply non-personalized recommendations based on average ratings, sex and type of movies. To make better predictions, we also improve our results based on the updated information and external data, such as IMDb and Rotten Tomatoes ratings.

4. Data Description

The dataset we get contains information about the name of classmates and their ratings for all the movies included in the survey. Also a DBMI data that contains similar information. However, the biggest problem is that not all movies are watched by us and everyone may watch different sets of movies.

To solve this problem and calibrate our collaborative filtering, for question 1, we go through the dataset and cover all the movies at least two of our team watched before. In detail, we choose 21 movies and 374 users from the dataset for filtering our model. After that, we choose 3 movies (Zero Dark Thirty, The Shape of Water, and Argo) we haven't watched before but are watched by many other classmates so that we could make recommendations. For finding the kindest and harshest member, we use all the movie data of our members and other classmates.

For question 2,3 and 4, besides the original dataset, we also scrape year, category of movies, classmates' gender and ratings from authoritative movie rating websites of each movie as an additional resource.

5. Model Development, Estimation and Results

5.1 How to predict your group members' ratings for the 3 movies of our choice?

As mentioned above, we choose Zero Dark Thirty, The Shape of Water, and Argo to make predictions. We use three different metrics - Cosine Similarity, Manhattan Distance and Euclidean Distance to make predictions. Since the dataset is not tremendous, Cosine similarity can predict the best.

Here is part of our predictive results:

Result 1: Cosine Similarity

	Zero Dark Thirty	The Shape of Water	Argo
Lifu	4	4	4
Yifan	4	4	5
Sungho	4	4	4
Xinyu	4	3	3

Result 2: Who's the kindest?

	Median	Average	Std.	Cosine(Proportion)
Lifu	5.0	4.5	0.8	0.2
Yifan	5.0	4.9	0.3	0.3
Sungho	4.5	4.5	0.6	0.3
Xinyu	4.0	3.6	0.6	0.2

We also want to find the kindest and harshest member of our group. To answer this question, as is shown in *Result 2,* we calculate the median, average scores and standard deviation of our members. Moreover, we made a virtual kindest classmate with all 5 ratings and compare the similarity between the virtual classmate and us. Based on the result, Yifan is the kindest and Xinyu is the harshest.

5.2 How to predict ratings of new movies?

There are many features that can describe a movie, and people may have similar ratings for specific movies. As a result, we collect information about the year of release, length of the movie and the movie category to make content-based prediction, and we use three methods to do the prediction.

The first method we use is to make predictions based on the category of movies. For example, the type of *Son of Saul* is war/drama. We select four most similar movies (like *Zero Dark Thirty*) that are the same type and have similar release year and length. Then we calculate the weighted average ratings of similar movies based on the ratings of our similar classmates/users and ourselves. The second way is based on the gender of our group members. For example, ladies may not be very interested in war-related movies, so their scores tend to be lower for war-related movies. We calculate the average of the same type pf movies based on gender and use it as the prediction value. For the third way, we find that some users in DBMI dataset have scores for all three movies. Thus, we deploy the user-user based *cosine* similarity matrix and compute the weighted average as our predicted value.

Result 3: Category-based

	Winter's Bone	*A Serious Man*	*Son of Saul*
Lifu	4	4	4
Yifan	4	5	5
Sungho	4	4	4

Result 4: gender-based

	Winter's Bone	*A Serious Man*	*Son of Saul*
Lifu	4	5	5
Yifan	4	5	5
Sungho	4	5	5

Result 5: DBMI-based

	Winter's Bone	*A Serious Man*	*Son of Saul*
Lifu	5	5	5
Yifan	5	5	5
Sungho	5	5	5

Xinyu	4	3	4	Xinyu	4	4	3	Xinyu	3	3	3

5.3 How to predict ratings of new audience?

Since lack of information about new customers, we predict their ratings based on non-personalized recommendation ways. We calculate the average ratings of total ratings, gender-based ratings and category-based ratings for three movies we required to predict.

Result 6: total data

	Avatar	The Wolf of Wall Street	Inception
Median	4	4	5
Average	4	4	4
Std.	1.02	0.95	0.85

Result 7: gender-based

	Avatar	The Wolf of Wall Street	Inception
Median	4	4	5
Average	4	4	5
Std.	0.84	0.76	0.66

Result 8: category-based

	Avatar	The Wolf of Wall Street	Inception
Median	4	4	4
Average	4	4	4
Std.	1.03	0.97	1.03

The standard deviation is stable, so we choose the average ratings to be our prediction ratings.

5.4 What will happen for predictive ratings of new audience with more information?

With the updated information, it becomes possible for us to use Cosine similarity for prediction. We try the z-scored cosine similarity between movies, z-scored cosine similarity between customers and use the average rating from our users, and the audience ratings from two movie authoritative websites to calculate user-to-website z-scored cosine similarity for prediction.

Result 9: total data

	Avatar	*The Wolf of Wall Street*	*Inception*
Shachi	3	3	4
Amy	4	4	5
Camille	3	3	4

Result 10: gender-based

	Avatar	*The Wolf of Wall Street*	*Inception*
Shachi	3	3	3
Amy	4	4	4
Camille	3	3	3

Result 11: authoritative websites-based

	Avatar	*The Wolf of Wall Street*	*Inception*
Shachi	2	3	3
Amy	4	4	3
Camille	3	3	1

Based on the comparison table (Appendix Page21), the prediction accuracy improved after involving new information. The improvement is from the updated customer-specific information, with which we can measure the similarity between different users, and the similarity between different movies.

6. Recommendations and Managerial Implications

Based on all of our analysis, it is worthwhile to develop a recommendation engine for any business that happens online. There are several reasons for this.

6.1 Increase total user-usage time and enjoy the compounded growth

A recommendation engine is crucial when it comes to dealing with an enormous pool of options. In the e-commerce retailing industry, such engines increase sales by showing a limited number of customized options to each customer. For video industry, we can expect the customers to watch more movies and stay longer on our websites after we get the recommendation engine running. In the long run, the benefit could be even larger as film copyrights do not expire. The recommendation system will be able to a growing pool of ‘inventories’ to leverage [1] and the firm can enjoy economies of scale. The effectiveness can be measured with a value function, which is introduced in Appendix Page 23 [2].

6.2 Enhance customer loyalty

As the recommendation engine becomes more accurate, the high-quality recommendation will potentially increase customer stickiness. Also, the firm can actually form customers’ movie tastes and watching habits through the recommendation system. Yet, this requires more detailed information which is collected based on the firm’s understanding of the customers. We encourage the firm to do customer analytics to better classify the customer body according to their demographic characteristics. The information can then be used to train the engine. Besides, the firm can also involve external sources of information to improve the accuracy of recommendations, especially for new movies.

7. Conclusion

In this report, we use different metrics to predict our group members’ ratings for 3 movies with 3 metrics under different situations. Although we are able to come up with reasonable predictions, there still exists uncertainty. For example, for the predictions of new audience, because of the lack of information, we can only predict the ratings using non-personalized recommendations. In the future, we can consider collecting more background information such as age group, ethnicity and preferences so that we may make sounder predictions. Also, after group members watch the movies and give the ratings, we can also test the accuracy of our filtering.

Reference

1. Anderson, Chris. "The Long Tail." Wired. Accessed November 28, 2022. https://www.wired.com/2004/10/tail/.
2. Sutton, R.S. and Barto, A.G. (no date) 'Reinforcement Learning: An Introduction', p. 352.
3. Leskovec, J., Rajaraman, A. and Ullman, J.D. (no date) 'Mining of Massive Datasets', p. 513.

Appendix

The formulas we use to calculate similarity between different classmates and users:

Cosine Similarity: $\frac{x \cdot y}{\|x\| * \|y\|}$

- $\|x\| = \sqrt{(x_a)^2 + (x_b)^2 + \cdots}$
- $\|y\| = \sqrt{(y_a)^2 + (y_b)^2 + \cdots}$
- $x \cdot y \; = \; x_a * y_a + x_b * y_b + x_c * y_c + \cdots$

Manhattan Distance: $|x_a - y_a| + |x_b - y_b| + \cdots$

Euclidean Distance: $\sqrt{(x_a - y_a)^2 + (x_b - y_b)^2 + \cdots}$

Q1:

Cosine similarity

1) Determine a set of movies

	[The King's Speech]	[La La Land]	[Inception]	[Avatar]	……	[Toy Story 3]	[The Imitation Game]
	MV1	MV2	MV3	MV4		MV20	MV21
AVERAGE	4.1136	3.9885	4.4904	4.0152		3.933	4.2377
STD	0.861	1.0615	0.8486	1.018		0.9375	0.8761
Lifu	5	5	4	5		3	
Yifan	5	5		5		5	5
Sungho		4	5				
Xinyu	4	3		3	3	3	4

2) Calibrate my filter on

[The King's Speech]	[La La Land]	[Inception]	[Avatar]	……	[Toy Story 3]	[The Imitation Game]
MV1	MV2	MV3	MV4		MV20	MV21
1.029474	0.952888	-0.57786	0.967472		-0.99524	0
1.029474	0.952888	0	0.967472		1.138097	0.870124

0	0.010828	0.600519	0		0	0
-0.13198	-0.93123	0	-0.99724		-0.99524	-0.27127

3-1) Choose any 3 movies

[Zero Dark Thirty]	[The Shape of Water]	[Argo]
CA1	CA2	CA3
3.898148	3.609756	3.921569
0.985311	1.091203	1.068786

3-2) Calibrate 3 movies

[Zero Dark Thirty]	[The Shape of Water]	[Argo]
CA1	CA2	CA3

4-1) Cosine Similarity(CA1)

	Sim(CA1,MV1)	Sim(CA1,MV2)	Sim(CA1,MV3)	Sim(CA1,MV20)	Sim(CA1,MV21)
Numerator	24.29954	13.85679	44.0357	25.45015	42.19558
Deno1	10.34408	10.34408	10.34408	10.34408	10.34408
Deno2	13.22876	16.12452	17.63519	14.93318	14.89966
Cosine	0.177577	0.083078	0.241398	0.164758	0.273778
AbsCosine	0.177577	0.083078	0.241398	0.164758	0.273778

Results:

	Zero Dark Thirty	*The Shape of Water*	*Argo*
Lifu	4.242771	4.12516	4.315408
Yifan	4.437734	4.235878	4.558022
Sungho	4.032134	3.681713	4.060337
Xinyu	3.50767	3.230797	3.498359

Manhattan Distance (Partial):

	Sim(CA1,MV1)	Sim(CA1,MV2)	Sim(CA1,MV3)	Sim(CA1,MV4)	Sim(CA1,MV5)

Total	183.6258	246.1892	259.6226	278.7068	120.1037
Lifu	1.029474	0.952888	0.577858	0.967472	0
Yifan	1.029474	0.952888	0	0.967472	0.885973
Sungho	0	0.010828	0.600519	0	0
Xinyu	0.131984	0.931231	0	0.99724	1.166891

5-2) Manhattan Distance(CA2)

	Sim(CA1,MV1)	Sim(CA1,MV2)	Sim(CA1,MV3)	Sim(CA1,MV4)	Sim(CA1,MV5)
Total	194.6514	237.1109	284.3454	281.8534	149.8342
Lifu	1.029474	0.952888	0.577858	0.967472	0
Yifan	1.029474	0.952888	0	0.967472	0.885973
Sungho	0	0.010828	0.600519	0	0
Xinyu	0.131984	0.931231	0	0.99724	1.166891

5-3) Manhattan Distance(CA3)

	Sim(CA1,MV1)	Sim(CA1,MV2)	Sim(CA1,MV3)	Sim(CA1,MV4)	Sim(CA1,MV5)
Total	176.2505	243.0302	264.1068	271.3306	114.9334
Lifu	1.029474	0.952888	0.577858	0.967472	0
Yifan	1.029474	0.952888	0	0.967472	0.885973
Sungho	0	0.010828	0.600519	0	0
Xinyu	0.131984	0.931231	0	0.99724	1.166891

Results:

	Zero Dark Thirty	*The Shape of Water*	*Argo*
Lifu	4.303799	4.054587	4.365538
Yifan	4.542811	4.305371	4.612538
Sungho	3.994612	3.727728	4.030541
Xinyu	3.459389	3.124961	3.442177

Euclidean Distance:

6-1) Euclidean Distance(CA1)

	Sim(CA1,MV1)	Sim(CA1,MV2)	Sim(CA1,MV3)	Sim(CA1,MV4)	Sim(CA1,MV5)
Total	15.27746	18.41973	18.16394	19.3454	12.15154
Lifu	1.059816	0.907995	0.33392	0.936002	0

Yifan	1.059816	0.907995	0	0.936002	0.784948
Sungho	0	0.000117	0.360623	0	0
Xinyu	0.01742	0.867192	0	0.994488	1.361635

6-2) Euclidean Distance(CA2)

	Sim(CA1,MV1)	Sim(CA1,MV2)	Sim(CA1,MV3)	Sim(CA1,MV4)	Sim(CA1,MV5)
Total	16.12595	17.67449	19.73952	18.97888	14.6557
Lifu	1.059816	0.907995	0.33392	0.936002	0
Yifan	1.059816	0.907995	0	0.936002	0.784948
Sungho	0	0.000117	0.360623	0	0
Xinyu	0.01742	0.867192	0	0.994488	1.361635

6-3) Euclidean Distance(CA3)

	Sim(CA1,MV1)	Sim(CA1,MV2)	Sim(CA1,MV3)	Sim(CA1,MV4)	Sim(CA1,MV5)
Total	15.16891	18.01405	18.60418	18.8652	12.16387
Lifu	1.059816	0.907995	0.33392	0.936002	0
Yifan	1.059816	0.907995	0	0.936002	0.784948
Sungho	0	0.000117	0.360623	0	0
Xinyu	0.01742	0.867192	0	0.994488	1.361635

Results:

	Zero Dark Thirty	*The Shape of Water*	*Argo*
Lifu	4.303799	4.054587	4.365538
Yifan	4.542811	4.305371	4.612538
Sungho	3.994612	3.727728	4.030541
Xinyu	3.459389	3.124961	3.442177

Who is the kindest?

- By average

			Raw Data				
Median	Average	Std.	User	[The Social Network]	[Amour]	[The King's Speech]	[La La Land]
5	4.5	0.785905248	Lifu			5	5
5	4.875	0.341565026	Yifan			5	5
4.5	4.5	0.577350269	Sungho				4
4	3.555555556	0.577350269	Xinyu	4	3	4	3

- By comparing with a virtual kindest man with all 5 ratings (partial)

Data

User	Lifu	Yifan	Sungho	Xinyu	Kindest	Kindest	Kindest	Kindest
[The Social Network]				4				5
[A Prophet]								
[Amour]				3				5
[The King's Speech]	5	5		4	5	5		5
[La La Land]	5	5	4	3	5	5	5	5

Results:

	Median	Average	Std.	Cosine(Proportion)
Lifu	5	4.5	0.785905248	0.248649
Yifan	5	4.875	0.341565026	0.251626
Sungho	4.5	4.5	0.577350269	0.250662
Xinyu	4	3.555555556	0.577350269	0.249063

Q2:

The calculations of gender-based analysis and the predictions based on DBMI dataset where someone has watched these movies are very similar with methods in Question 1 and Question 2. In this part, we will mainly cover the calculation of categoric-based predictions.

The movie type, length and release of year information we collect:

Movie Name	type	year	length
[The Social Network]	Drama/History	2010	2
[A Prophet]	Drama/Crime	2009	2.35
[Amour]	Romance/Drama	2012	2.07

[The King's Speech]	Drama/History	2010	1.58
[La La Land]	Musical/Romance	2016	2.08
[Boyhood]	Drama/Coming-of-age story	2014	2.43
[Inception]	Action/Sci-fi	2010	2.28
[A Separation]	Drama/Mystery	2011	2.03
[The Artist]	Drama/Romance	2011	1.4
[The White Ribbon]	Drama/War	2009	2.24
[Zero Dark Thirty]	War/Thriller	2012	2.37
[Avatar]	Sci-fi/Adventure	2009	2.42
[Spotlight]	Drama/Indie film	2015	2.08
[Precious]	Drama/Indie film	2009	1.5
[The Tree of Life]	Drama/Experimental	2011	2.18
[12 Years a Slave]	Drama/History	2013	2.14
[Blue is the Warmest Colour]	Romance/Drama	2013	3
[Son of Saul]	War/Drama	2015	1.47
[Up in the Air]	Romance/Drama	2009	1.49
[Inglourious Basterds]	War/Action	2009	2.33
[Mad Max: Fury Road]	Action/Adventure	2015	2
[Moonlight]	Drama/Indie film	2016	1.51
[Birdman]	Drama/Comedy	2014	2
[Manchester by the Sea]	Drama	2016	2.17
[Lincoln]	War/Drama	2012	2.3
[Hugo]	Fantasy/Adventure	2011	2.06
[Toni Erdmann]	Drama/Comedy	2016	2.42
[The Shape of Water]	Romance/Fantasy	2017	2.03
[Three Billboards Outside Ebbing, Missouri]	Crime/Drama	2017	1.55
[Argo]	Thriller/Drama	2012	2
[Gravity]	Sci-fi/Thriller	2013	1.31
[Black Swan]	Drama/Thriller	2010	1.5
[Ida]	Drama	2013	1.22
[Leviathan]	Sci-fi/Horror	1989	1.38
[The Wolf of Wall Street]	Comedy/Drama	2013	3
[True Grit]	Western/Drama	2010	1.5
[The Descendants]	Drama/Comedy drama	2011	1.55
[The Secret in Their Eyes]	Thriller/Mystery	2015	1.51
[Life of Pi]	Adventure/Drama	2012	2.07
[Arrival]	Sci-fi/Thriller	2016	1.56
[Call Me by Your Name]	Romance/Drama	2017	2.1
[Winter's Bone]	Drama/Drama	2010	1.4
[The Grand Budapest Hotel]	Comedy/Drama	2014	1.4

[Dunkirk]	War/Action	2017	1.46
[Inside Llewyn Davis]	Drama/Music	2013	1.45
[A Serious Man]	Comedy/Drama	2009	1.45
[Toy Story 3]	Adventure/Family	2010	1.43
[Beasts of the Southern Wild]	Drama/Fantasy	2012	1.31
[The Imitation Game]	War/Drama	2014	1.54
[The Fighter]	Drama/Sport	2010	1.55

The classmate gendor information we collect:

Serial Number	Gender	First name	Last name
1	F	Teagan	Towhey
2	F	Anushka	Shah
3	M	Sanka Naga Nitesh	Nitesh
4	M	Akash	Puthalath
5	M	Arjun	Mahesh
6	F	Parina	Kolhe
7	F	Vindhya	Mandekar
8	M	Ridwan	Abduslaam
9	M	Issam "Sam"	Tamer
10	F	Haridhakshini	SubramoniaPillai Ajeetha
11	M	Venkata Aravind Sampath	Bhagavatula
12	M	Anirudh	Menon
13	F	Mrinalini Sri	Dosapati
14	F	Aishwarya Prashant	Kamat
15	M	Laksh	Suryanarayanan
16	F	Sanjana	Kallol
17	F	Sheetal	Rajgure
18	F	Yashi	Tiwari
19	F	Pradeepthi	Mallappa
20	F	Abhigna	Anilkumar
21	F	Alba	Valdivia Plummer
22	M	Karan	Thakkar
23	M	Xu	Zhang
24	F	Aakriti	Pande
25	F	Sneha	Guravannavar
26	F	Anisha	Samant
27	F	Jyothika	Mohan
28	F	Kexin (Shera)	Huang
29	M	Jake	Brophy
30	M	Chi En	Hwang

31	M	Murad	Salamov
32	M	Hanish	Singla
33	F	Shiyi	Yue
34	M	Wenbo	Wang
35	F	Anchal	Chaudhary
36	M	Varenium	Setia
37	M	Qiuhong	Wei
38	M	Oscar	Xu
39	F	Qian	Wu
40	F	Meng-Wei	Wu
41	M	Zhihan (Jimmy)	Jiang
42	M	Zheming (Jim)	Xu
43	F	Ching-Wen	Huang
44	M	Siwei	Ran
45	F	Siqi	Chen
46	F	Blessia	Li
47	F	Shirley	Deng
48	F	Jialu	Wang
49	F	Yuwen	Ma
50	F	Yongxin	Lin
51	F	Yifan	CAI
52	M	Liam	Wan
53	M	Shaolong	Xue
54	F	Tamalika	Basu
55	M	Trishal	Jadhav
56	F	Sahithi	Sukhavasi
57	F	Xingyi (Stella)	Wang
58	F	Anandita	Juneja
59	M	Rahul	Rajput
60	M	Domenic	Diaa
61	F	Priyanka	Murugan
62	F	Aishwarya	.
63	M	Ajaiy Praveen	Thiruchelvam
64	M	Raghav Rama	Bhadran
65	F	Sripriya	Srinivasan
66	F	Arpita	Mangal
67	M	Andrew	Hamaty
68	F	Vaaridhi	Mathur
69	F	namuun	boldbaatar
70	F	SRISHTI	AGARWAL

71	F	Priya	Iddalgi
72	M	Sumanth	Munnangi
73	F	Chen	Zhengjia
74	M	anant	bairagi
75	M	Rohith Reddy	Amanaganti
76	M	Vamsee Krishna Reddy	Narahari
77	F	Hanqiu	Yu
78	F	Rishika	Chaudhary
79	F	Alice	Shen
80	M	Sayar	Banerjee
81	F	Xianzhang	Deng
82	M	Lifu	Ma
83	F	Jinny	Zhong
84	F	Zhiyu	Zhang
85	M	Kshitij	Karan
86	M	Yuto	Takeda
87	M	Yuequn	Yu
88	M	Yifan	Zhu
89	M	Kangjian	Gao
90	F	Jiayi	Jiang
91	F	Suzana	Amer
92	M	Sungho	Lee
93	M	Daoxin	Wang
94	F	Xinyu	Liu
95	M	Arjun	Remeshkumar Nair
96	F	Harshkriti	Kaur
97	F	Karina	Munjal
98	M	Shulang	Ning

Step 1: use Cosine similarity to find the most similar classmates/users of everyone

TOP SIMILAR USERS of MARIO

Sanka Naga Nitesh	0.55101042
Vindhya	0.61149905
Aakriti	0.48790223
User 20	0.51665607
User 180	0.48549794

TOP SIMILAR USERS of Yifan

Yuequn	0.72474861
User 10	0.74942713
User 56	0.6428317
User 161	0.4923336
User 220	0.48337799

TOP SIMILAR USERS of Sungho

Sneha	0.63900965
Harshkriti	0.6761234
User 172	0.67082039
User 231	0.65483077
User 262	0.625

TOP SIMILAR USERS of Xinyu

Kexin (Shera)	0.55448321
Rishika	0.52650052
User 26	0.55743859
User 86	0.49628248
User 269	0.5084323

Step 2: Find similar movies and make predictions

	Lifu	Yifan	Sungho	Xinyu
[Winter's Bone]	3.85183066	4	4	4.25713556
Drama/Drama				
2010				
Length: 1.4				
[The Social Network]	3.09824157	3	4	4.51427113
[A Prophet]	3	4		
[The Artist]	2.54826556			
[Up in the Air]	2.16813818	5	4	

	Lifu	Yifan	Sungho	Xinyu
[A Serious Man]	4.48251437	4.68245072	3.95752578	3.20053646
Comedy/Drama				
2009				
Length: 1.45				
birdman	3.63583159	4	3.55304021	

[Toni Erdmann]	5	4		
[The Wolf of Wall Street]	4.22428333	4.14205652	4.31953714	3.60160938
[The Descendants]	3		4	

	Lifu	Yifan	Sungho	Xinyu
[Son of Saul]	4.33592731	4.63887205	4.162781938	4
War/Drama				
2015				
Length: 1.47				
[The White Ribbon]	4		3.488345814	4
[Inglourious Basterds]		4.20346411	4	
[Lincoln]				
[The Imitation Game]	4.67185462	3.62976822	5	4

Step 3: Final results

Final Prediction

	[Winter's Bone]	Round	[A Serious Man]	Round	[Son of Saul]	Round
Lifu	3.851830664	4	4.482514365	4	4.335927309	4
Yifan	4	4	4.682450724	5	4.638872054	5
Sungho	4	4	3.957525784	4	4.162781938	4
Xinyu	4.257135565	4	3.200536459	3	4	4

Q3: predict for new customers

1.Direct average

User	[Avatar]	[The Wolf of Wall Street]	[Inception]
Median	4	4	5
Average	4.0151515	4.11683849	4.4903846
Std.	1.0179607	0.95054883	0.8486251
Lifu	5	5	4
Yifan	5	5	
Sungho			5
Xinyu	3	3	

2.Female average

User	[Avatar]	[The Wolf of Wall Street]	[Inception]
Median	4	4	5
Average	4.133333	4.166667	4.568182
Std.	0.842075	0.762431	0.661138
Xinyu	3	3	

3-1. Category (Sci-fi for Avatar, Inception)

3-2. Category (Comedy/Drama for The Wolf of Wall Street)

User	User	[The Wolf of Wall Street]
Median	Median	4
Average	Average	4.042105
Std.	Std.	0.966633

	Avatar	*The Wolf of Wall Street*	*Inception*
Median	4	4	5
Average	4.0151515	4.11683849	4.4903846
Std.	1.0179607	0.95054883	0.8486251

	Avatar	*The Wolf of Wall Street*	*Inception*
Median	4	4	5
Average	4.133333	4.166667	4.568182
Std.	0.842075	0.762431	0.661138

	Avatar	*The Wolf of Wall Street*	*Inception*
Median	4	4	4
Average	4.125867	4.042105	4.125867
Std.	1.029821	0.966633	1.029821

Q4:

So far, our analysis uses only internal data. To further test whether including external data would improve accuracy, we experimented with scraped movie ratings from IMDb and Rotten Tomatoes. We included the average rating from our customers, and the audience ratings from the two websites and used user-to-website z-scored cosine similarity for prediction.

Here are the rating collected from authoritative movie websites: IMDb and Rotten Tomatoes data (source: IMDb, Rotten Tomatoes; as of November 2022)

Name	*IMDb Rating (out of 10)*	*IMDb Ranking*	*RT Ratings (out of 100)*	*RT Ranking*
The Social Network	7.8	21	87	13
A Prophet	7.8	21	89	10
Amour	7.9	16	82	24
The King's Speech	8	12	92	3
La La Land	8	12	81	27
Boyhood	7.9	16	80	30
Inception	8.8	1	91	5
A Seperation	8.3	2	92	3
The Artist	7.9	16	87	13
The White Ribbon	7.8	21	79	35
Zero Dark Thirty	7.4	35	80	30
Avatar	7.8	21	82	24
Spotlight	8.1	7	93	1
Precious	7.3	41	81	27
The Tree of Life	6.8	49	60	50
12 Years a Slave	8.1	7	90	7

La vie d'Adèle	7.7	29	85	19
Saul fia	7.4	35	80	30
Up in the Air	7.4	35	79	35
Inglourious Basterds	8.3	2	88	12
Mad Max: Fury Road	8.1	7	86	16
Moonlight	7.4	35	79	35
Birdman	7.7	29	78	41
Manchester by the Sea	7.8	21	78	41
Lincoln	7.3	41	80	30
Hugo	7.5	34	78	41
Toni Erdmann	7.3	41	73	47
The Shape of Water	7.3	41	72	48
Three Billboards Outside				
Ebbing, Missouri	8.1	7	87	13
Argo	7.7	29	90	7
Gravity	7.7	29	79	35
Black Swan	8	12	84	21
Ida	7.4	35	79	35
Leviathan	5.8	50	80	30
The Wolf of Wall Street	8.2	5	83	23
True Grit	7.6	33	85	19
The Descendants	7.3	41	79	35
El secreto de sus ojos	8.2	5	93	1
Life of Pi	7.9	16	84	21
Arrival	7.9	16	82	24

Call Me by Your Name	7.8	21	86	16
Winter's Bone	7.1	47	76	44
The Grand Budapest Hotel	8.1	7	86	16
Dunkirk	7.8	21	81	27
Inside Llewyn Davis	7.4	35	74	46
A Serious Man	7	48	68	49
Toy Story 3	8.3	2	90	7
Beasts of the Southern Wild	7.2	46	76	44
The Imitation Game	8	12	91	5
The Fighter	7.8	21	89	10

We also use three ways to do the collaborative filtering. The calculation process is similar with Question 1, so we skip the detailed calculation here.

Results:

	Avatar	*The Wolf of Wall Street*	*Inception*
Shachi	2.371990	2.766965	3.314986
Amy	4.323294	3.922960	2.720969
Camille	3.056155	2.593082	1.487061

We noticed that for users with higher cosine similarity, the prediction accuracy did not improve from the prediction with only internal data, even for customers with higher similarity with the website ratings.

The comparison between results of Q3 and Q4:

Shachi

	Precious	*12 Years a Slave*	*Mad Max: Fury Road*	*Black Swan*	*Toy Story 3*	*Avg*

Real Rating	2.00	2.00	4.00	3.00	3.00	2.80
Previous Prediction	3.72	4.04	3.62	4.12	3.93	3.89
| Real - Previous |	1.72	2.04	0.38	1.12	0.93	**1.24**
Updated Prediction	2.72	2.54	3.60	2.84	2.80	2.89
| Real - Updated |	0.72	0.54	0.40	0.16	0.20	**0.41**

Amy

	Precious	*12 Years a Slave*	*Mad Max: Fury Road*	*Black Swan*	*Toy Story 3*	*Avg*
Real Rating	4.00	5.00	5.00	4.00	3.00	4.20
Previous Prediction	3.72	4.04	3.62	4.12	3.93	3.89
| Real - Previous |	0.28	0.96	1.38	0.12	0.93	**0.73**
Updated Prediction	4.20	4.40	4.73	4.10	3.57	4.20
| Real - Updated |	0.20	0.60	0.27	0.10	0.57	**0.35**

Camille

	Precious	*12 Years a Slave*	*Mad Max: Fury Road*	*Black Swan*	*Toy Story 3*	*Avg*
Real Rating	4.00	3.00	4.00	1.00	4.00	3.20
Previous Prediction	3.72	4.04	3.62	4.12	3.93	3.89
| Real - Previous |	0.28	1.04	0.38	3.12	0.67	**0.98**
Updated Prediction	3.28	3.25	3.77	2.27	3.65	3.24
| Real - Updated |	0.72	0.25	0.23	1.27	0.35	**0.56**

The result shows the absolute difference between prediction and real ratings from new customers. The accuracy of prediction improved. The improvement is from the updated customer-specific information, with which we can measure the similarity between the new customers with other existed users, and the similarity between the predicted movies and rated movies.

Value Function for measuring effectiveness of the recommendation system:

$$Q_\pi(s_t, a_t) = \mathbb{E}[U_t | S_t = s_t, A_t = a_t]$$

which measures the return of recommending a movie (taking action a) at the time t given the customer's current ratings (situation s) following a recommendation policyπ.