Movie Recommendation System

Our Team:(Team Outliers)
Sanket J Shah AU1841120
Kishan Mehta AU1841072
Harsh P Patel AU1841075
Dhruv Sharma AU1841102

Introduction

- With the rapidly growing Internet, The content made available everyday can easily make an average user **overwhelmed.**
- Even during the Pandemic OTT platforms in India have received **80% growth** in new users, let alone the world.
- The emergence of the online media sharing sites (Netflix, Hulu or even YouTube) have faced **new challenges in content recommendation.**
- Recommended System collects data from users activities and analyzes the data to generate **customized recommendations.**

Problem Statement

A recommendation system that takes into account all the required aspects of a movie and make relevant recommendations.

GANTT CHART

	Week 1	Week 2	Week 3	Week 4	Week 5	Week 6	Week 7	Week 8	Week 9
Reasearch(literature review) on Recomendation System									
Research(literature review) on Content based system and consine similarity									
Coding(and data collection) of Cosine Similarity									
Research(literature review) on Collaborative Filtering system									
MidSem documentation									
Research(literature review) on Gaussian Mixture Model (GMM)									
Coding(and data collection) of GMM with scikit learn									
Coding of GMM without scikit learn									
Final documentation									

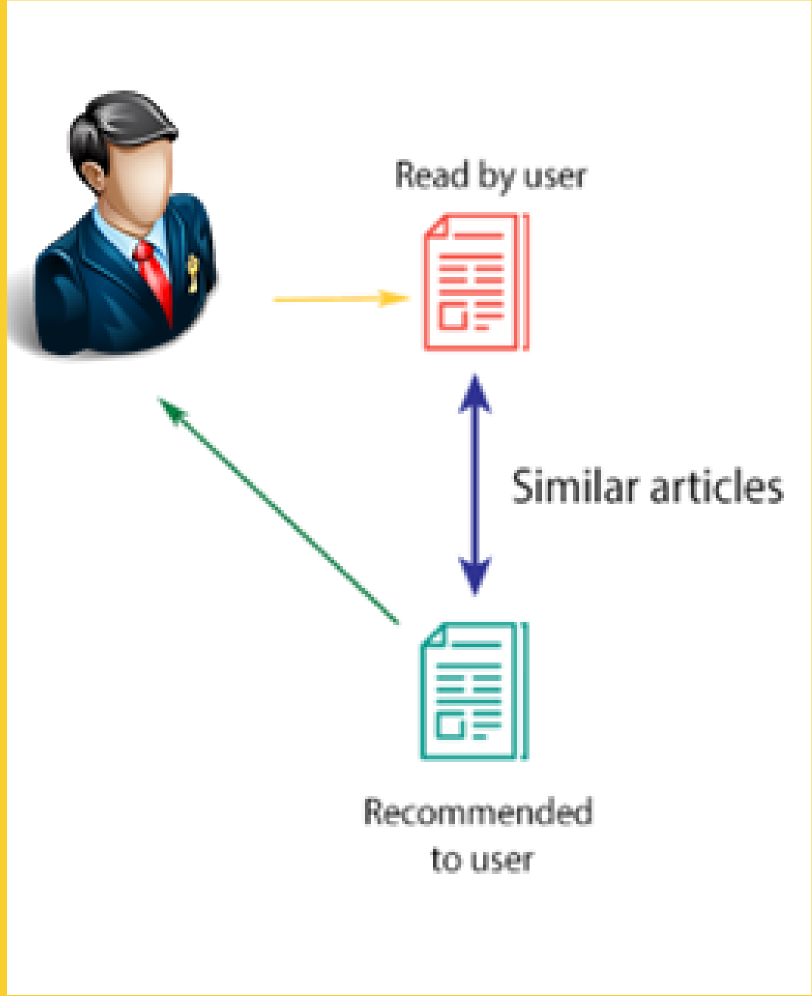
Existing Body of Work

- Flickmetrix recommends movies based on their availability on OTT platforms and it takes many inputs from user while searching for movie such as release year and rating.[1]
- Date Night lets you add two movies to the app and it will spit out a line of recommendations that are somewhere between the two choices.[2]
- Movie of the Night recommendation works just like Flickmetrix by taking in various tedious inputs from users but this recommendation system only suggests one movie based on the criteria.[3]

Existing Body of Work

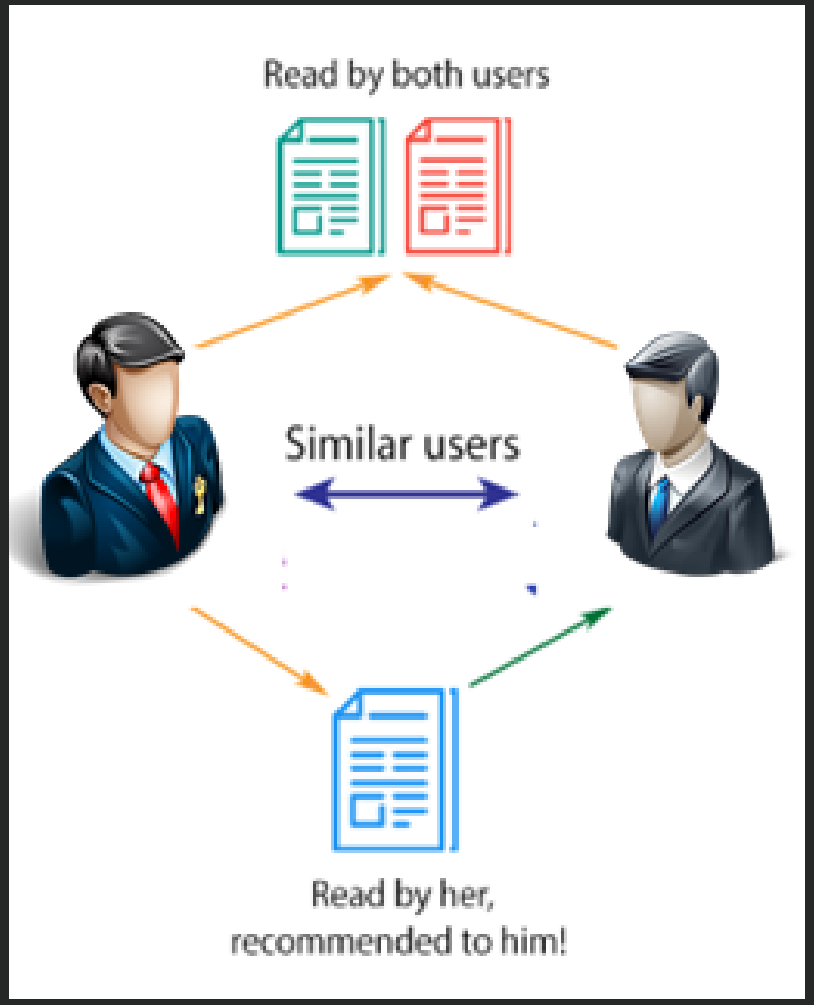
Content Based Filtering

- It is a technique that recommends the user some content on the bases of previous preferences.



Collaborative Filtering

- It is a technique that can filter out items that a user might like on the basis of other user's liking.



Our Approach

Content Based Filtering

TF-IDF based cosine similarity approach.
Dataset: MovieLens 25M Dataset[4]

TF-IDF: Convert string input to numeric output

Cosine Similarity: Cosine similarity is used to calculate the similarity between two vector values.

Collaborative Filtering

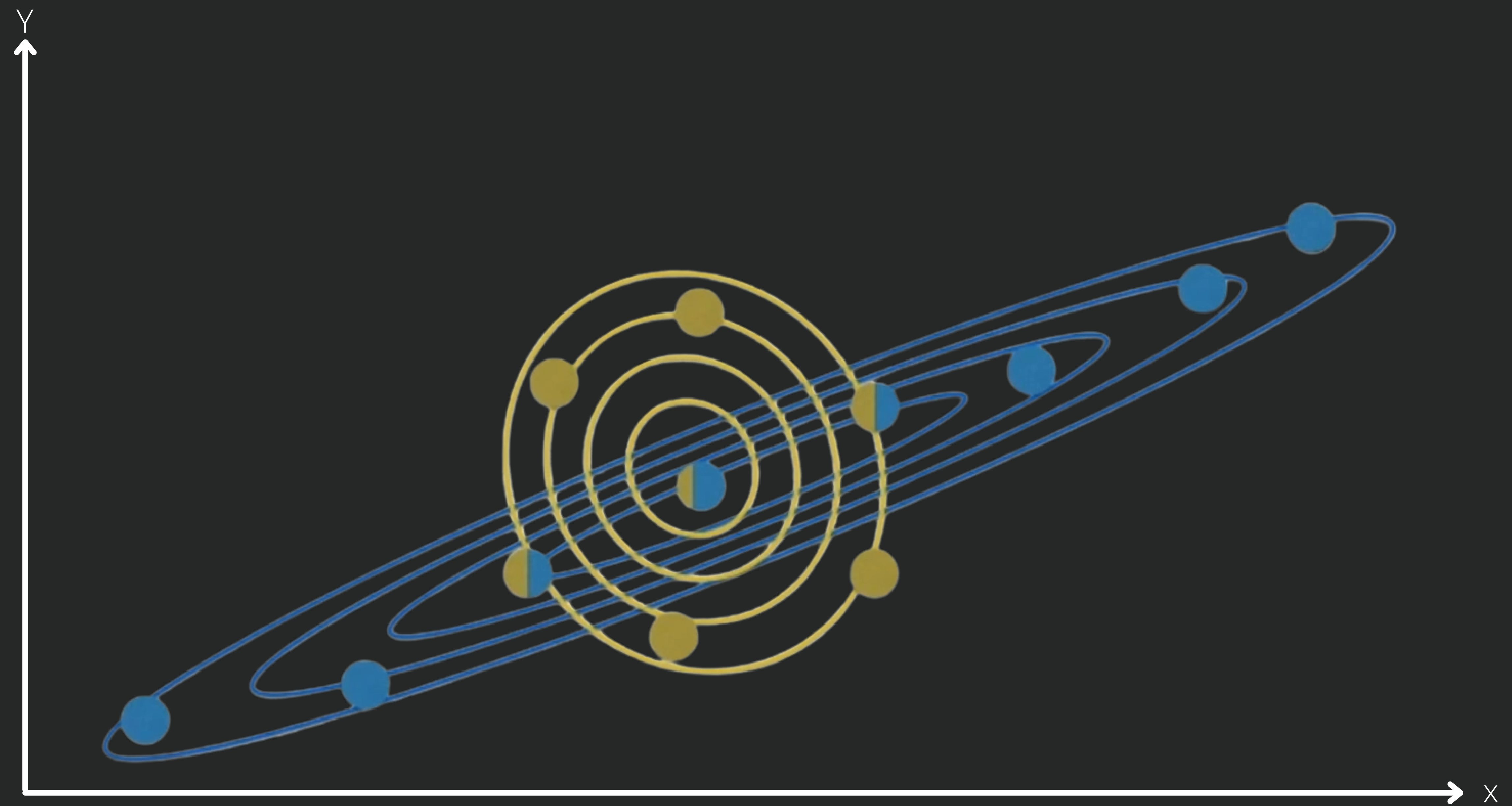
User Clustering using Gaussian Mixture Model with Expectation Maximization
Dataset: MovieLens 100K Dataset[5]

Collaborative Filtering: Predicting the rating of the movie based on rating of other users.
GMM: Well known technique of soft clustering. Uses linearly added Gaussian distrubutions to form the probability density function.
EM: An algorithm to map clusters on the unsupervised data points in GMM.

Approach to implement GMM with EM

1. Extracting User and Movie Data from the 100K dataset
2. Cleaning Data and making it readable
3. Creating a Rating Matrix
4. Create the GMM with specified number of clusters
5. Fit the model on our Rating Matrix using the E & M steps
6. Calculate responsibilities once again after the max iterations end
7. Use these responsibilities to predict the user's ratings

Responsibility Visualization



Final Results

- Cosine Similarity using Movie Metdata and IMDB's weighted rating formulae

```
In [62]: ImprovedRecommendations('The Dark Knight')
```

Out[62]:

	title	vote_count	vote_average	year	wr
7648	Inception	14075	8	2010	7.919065
8613	Interstellar	11187	8	2014	7.898936
6623	The Prestige	4510	8	2006	7.762198
3381	Memento	4168	8	2000	7.744491
8031	The Dark Knight Rises	9263	7	2012	6.922734
6218	Batman Begins	7511	7	2005	6.905676
1134	Batman Returns	1706	6	1992	5.848168
132	Batman Forever	1529	5	1995	5.051917
9024	Batman v Superman: Dawn of Justice	7189	5	2016	5.013324
1260	Batman & Robin	1447	4	1997	4.281221

Final Results

- Gaussian Mixture Model clustering using Expectation Maximization

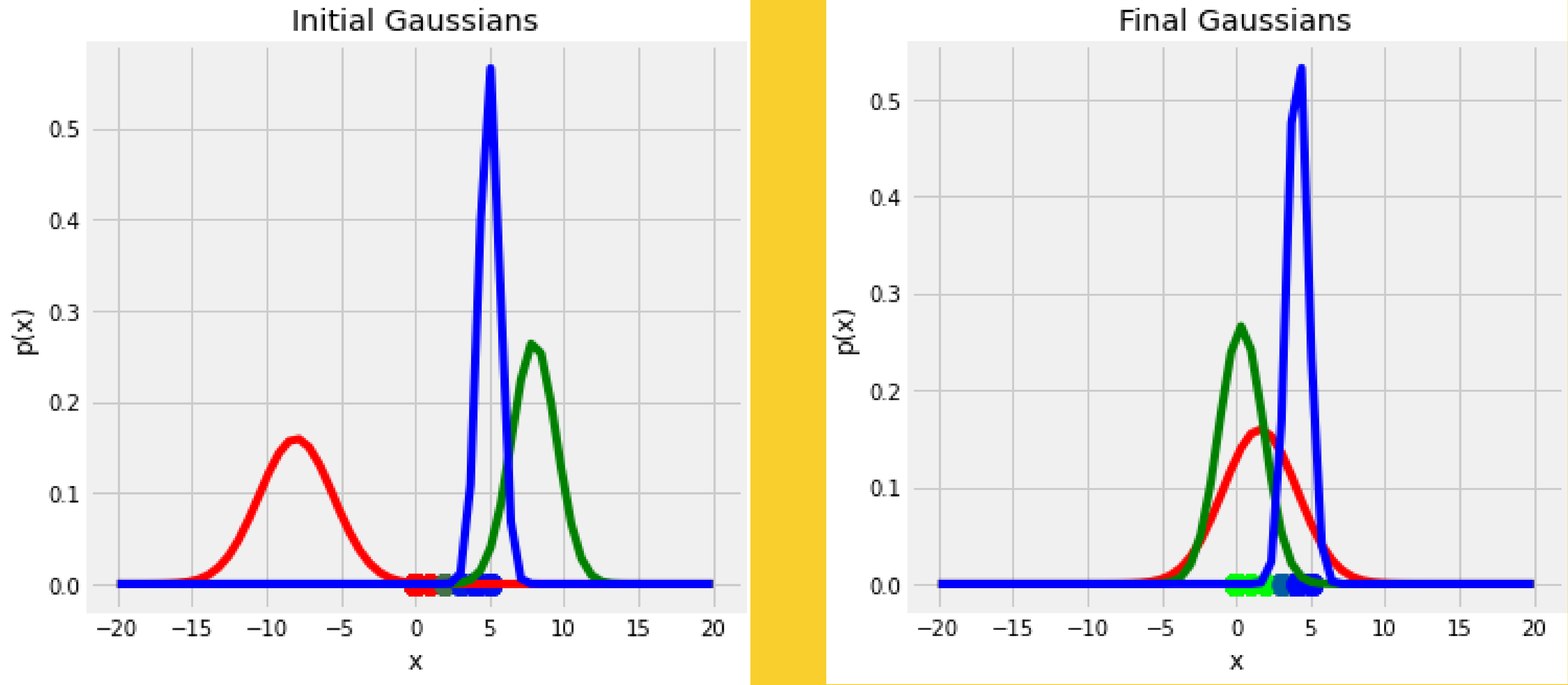
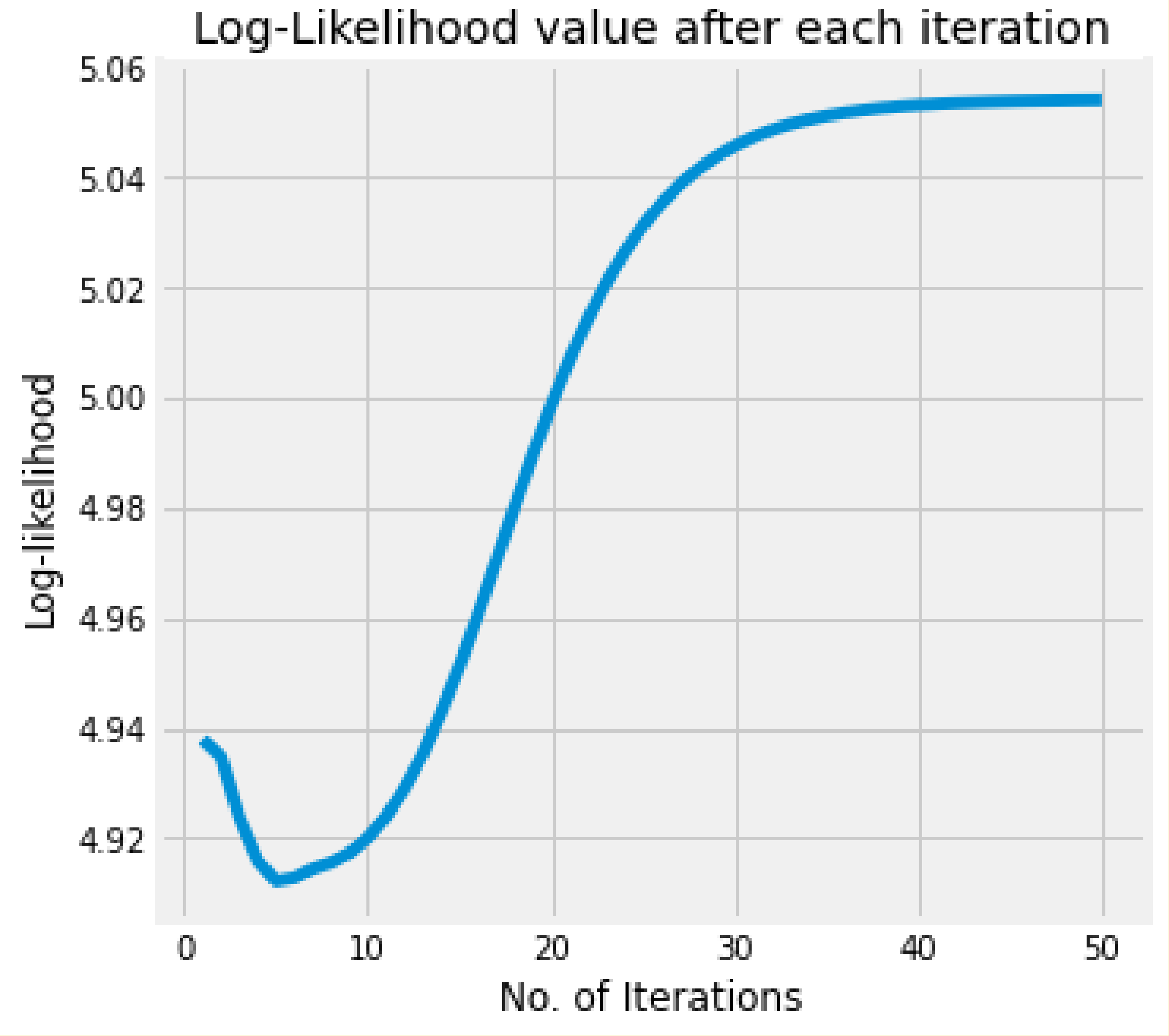


Fig. Visualization of distributions fitting to the datapoints using EM

Final Results

- The log likelihood of EM steps with increasing iterations



Final Results

- The original matrix and the predicted matrix with 2 gaussian distributions

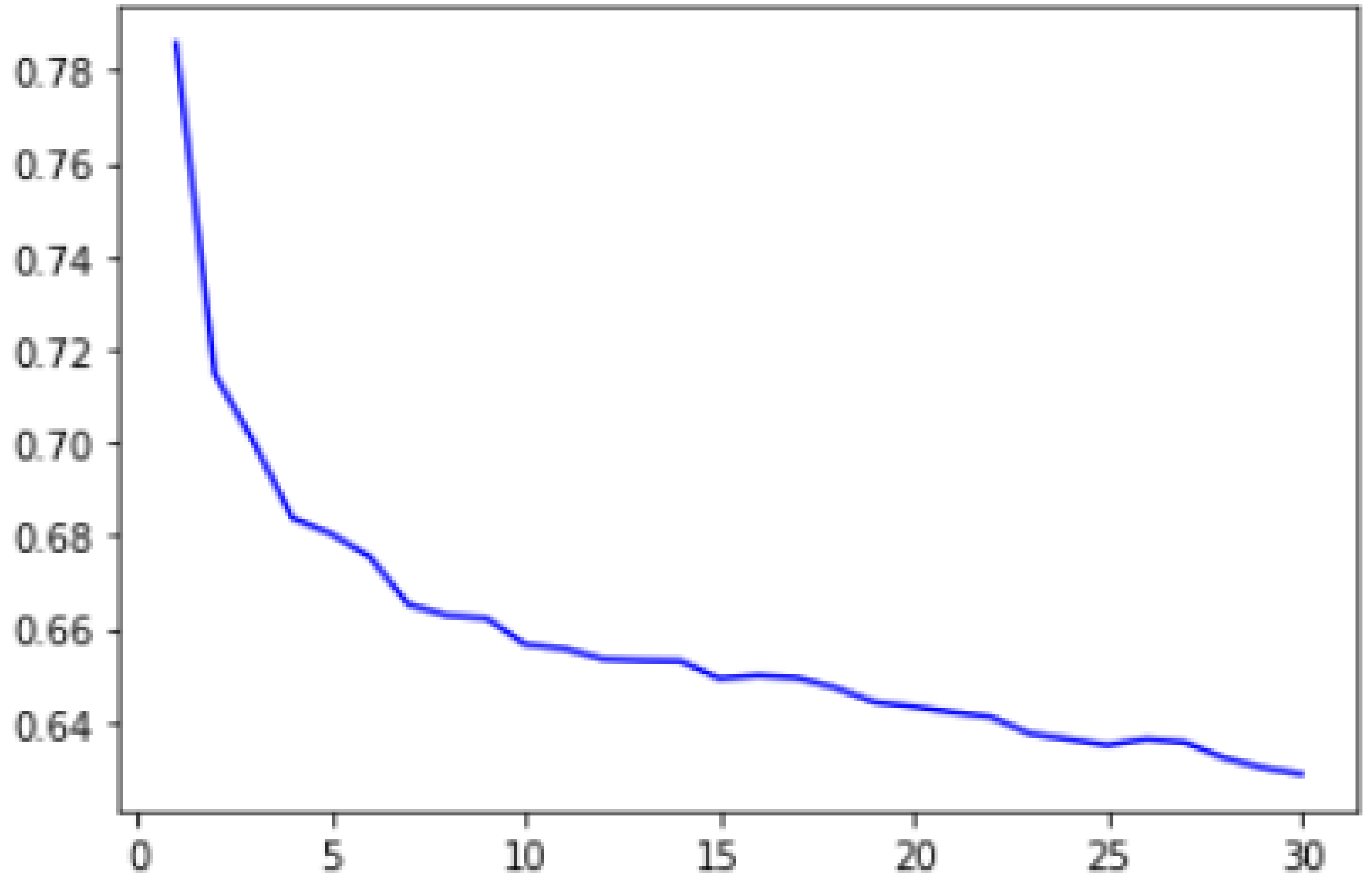
```
Original Matrix:
[[5. 3. 4. ... 0. 0. 0.]
 [4. 0. 0. ... 0. 0. 0.]
 [0. 0. 0. ... 0. 0. 0.]
 ...
 [5. 0. 0. ... 0. 0. 0.]
 [0. 0. 0. ... 0. 0. 0.]
 [0. 5. 0. ... 0. 0. 0.]]
Predicted Matrix:
[[5. 3. 4. ... 0. 0. 0.]
 [4. 0. 0. ... 0. 0. 0.]
 [1. 0. 0. ... 0. 0. 0.]
 ...
 [5. 0. 0. ... 0. 0. 0.]
 [1. 0. 0. ... 0. 0. 0.]
 [3. 5. 1. ... 0. 0. 0.]]
0.7147657346241669
```

Final Results

- The RMSE for the predicted ratings for different number of dsitributions

```
array([0.        , 0.7857215 , 0.71476573, 0.70016537, 0.68379831,
       0.68040474, 0.67541847, 0.66518173, 0.66286884, 0.66229983,
       0.65664451, 0.65576251, 0.65355135, 0.65324355, 0.65311372,
       0.64932882, 0.64998437, 0.6494725 , 0.64732167, 0.64423562,
       0.64327536, 0.64203533, 0.64102801, 0.63749579, 0.63630593,
       0.6350518 , 0.63632723, 0.63563975, 0.63225194, 0.63015889,
       0.62883185])
```

RMSE



Number of distributions

Conclusion

COSINE SIMILARITY APPROACH:

The results of this approach were satisfying but they are same for all the users (who have watched the same movie)

GMM WITH EM APPROACH:

The results of this approach were logically inline with collaborative filtering but were not practically applicable.

FINAL THOUGHTS

A hybrid system that uses a combination of both the approaches would be practical and give best results in terms of user satisfaction.

ROLE OF EACH MEMBER

Sanket
- Coding using scikit
- Literature review
- Data Cleaning
- Documentation

Harsh
- Coding without scikit
- Literature review
- Data cleaning
- Documentation

Kishan
- Literature review
- Data visualizing
- Report writing
- Documentation

Dhruv
- Literature review
- Data gathering
- Report writing
- Documentation

References

1. Flickmetrix- Link
2. Date Night- Link
3. Movie of the Night- Link
4. Dataset 25m- Link
5. Dataset 100k- Link

Thank You