EE219 Project 3 - Report
Collaborative Filtering
Winter 2018

Introduction

Collaborative filtering, also referred to as social filtering, filters information by using the recommendations of other people. It is based on the idea that people who agreed in their evaluation of certain items in the past are likely to agree again in the future.

In this project, we explore two types of collaborative filtering methods:

1. **Neighborhood-based collaborative filtering:** In the neighborhood-based approach, a number of users are selected based on their similarity to the active user. A prediction for the active user is made by calculating a weighted average of the ratings of the selected users. In this project we have made use of the Pearson coefficient as a measure of correlation to find the similarity between users and k-nearest neighbours algorithm for collaborative filtering.
2. **Model-based collaborative filtering:** In the model based approach, models are developed to predict user's rating of unrated items. In this project, we explore the latent factor based models for collaborative filtering.

Dataset

The MovieLens dataset is used in this project. The dataset consists of ratings of 671 users for 9066 movies.

Q1: Sparsity of the movie rating dataset

The **sparsity of the dataset** is the fraction of the user/item rating matrix that is not empty. It is one minus the ratio of available ratings to possible ratings. The sparsity of the dataset was found to be **0.9836671498274911**.

Q2: Histogram showing frequency of the rating values

A histogram for the frequency of different ratings was plotted. Most of the ratings are in the 3-5 range . The proportion of low ratings in the dataset is small.

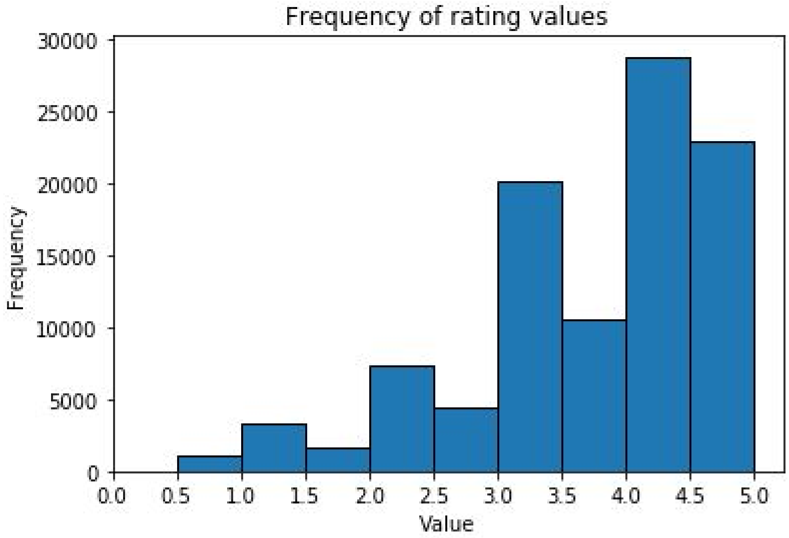


Figure 1: The frequency of ratings

Q3: Plot showing distribution of ratings among movies

The frequency distribution of the number of ratings for different movies was plotted. The most widely rated movie has 271 ratings. The least rated movie has 1 rating.

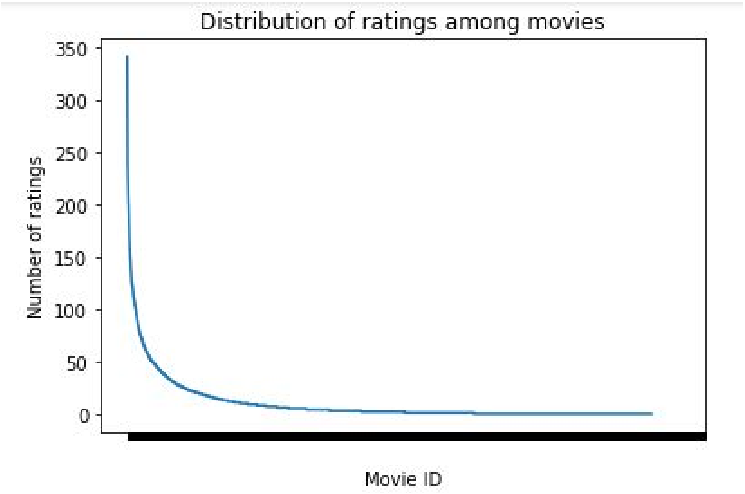


Figure 2: The distribution of number of ratings for a movie

X labels (Movie_ids ordered by decreasing frequency): [356, 296, 318, 593, 260, 480, 2571, 1, 527, 589, 1196, 110, 1270, 608, 2858, 1198, 780, 1210, 588, 457, 590, 2959, 47, 50, 150, 364, 858, 4993, 380, 592, 32, 2762, 2028, 1580, 5952, 377, 595, 7153, 344, 4306,.. .. .. .., 143859, 158314, 1692, 3845, 3909, 167, 563, 127124, 134246, 134528, 134783, 137595, 138204, 60832, 64997, 72380, 129, 4736, 6425]

Q4: Plot showing distribution of ratings among users

The frequency distribution of the number of ratings by each user was plotted. The maximum number of ratings of a user was found to be 2391 and minimum number of ratings 20.

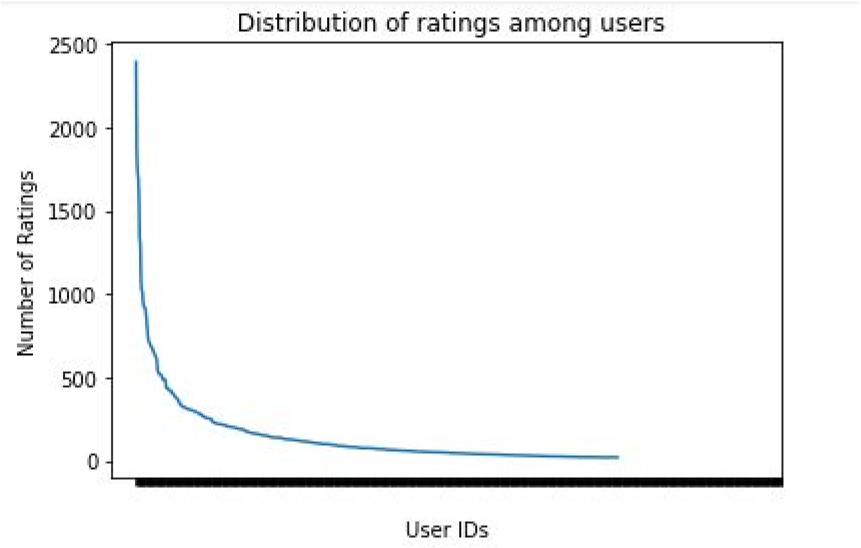


Figure 3: The distribution of the number of ratings by a user

X labels (User IDs ordered by decreasing frequency): [547, 564, 624, 15, 73, 452, 468, 380, 311, 30, 294, 509, 580, 213, 212, 472, 388, 23, 457, 518, 461, 232, 102, 262, 475, 306, 119, 654, 358, 529, 575, 105, 56, 353, 664, 48, 587, 165, 596, 195, 384, 463, 605, ..., 347, 413, 459, 490, 513, 549, 566, 634, 635, 24, 45, 64, 112, 127, 158, 174, 276, 356, 368, 469, 477, 506, 579, 1, 14, 35, 76, 209, 221, 249, 289, 296, 310, 319, 325, 337, 399, 438, 444, 445, 448, 484, 485, 498, 540, 583, 604, 638, 651, 657, 668]

Q5: Explain the salient features of the distribution found in question 3 and their implications for the recommendation process.

The number of ratings received by a movie is a measure of its popularity. The number of ratings per movie in the dataset ranges from 1-271. The popular movies can introduce a bias in the recommendation systems i.e popular movies are more likely to be recommended. It would be useful for a recommendation system to make use of weighting functions that are used in measuring users' similarities so that the effect of popular items is decreased with similar opinions and increased with dissimilar ones.

Q6: Variance of the rating values received by each movie

The variances in the ratings of different movies was calculated and a histogram capturing the frequencies of different variance values was plotted.

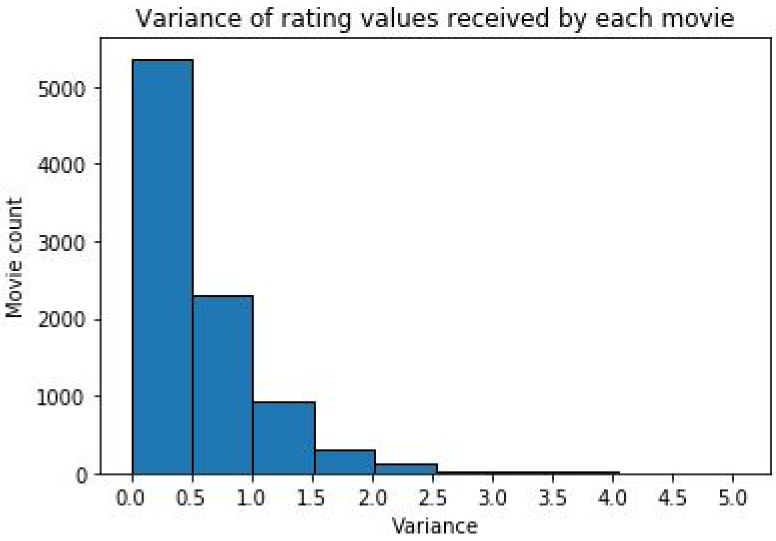


Figure 4: Histogram of variances in ratings

A large number of the movies have a low variance(0-0.5) among their ratings which means that most of the movies receive the same ratings by its viewers. The majority of the movies have a variance in their rating which is less than 2.5, showing that most movies receive ratings that are fairly similar.

Q7: Write down the formula for μ_u in terms of I_u and r_{uk}

$\mu_u = \Sigma (r_{uk})/len(I_u)$ for all k belonging to Iu

Q8: In plain words, explain the meaning of $I_u \cap I_v$. Can $I_u \cap I_v = \Phi$?

Iu ∩ Iv is the set of movies rated by users u and v. Yes,the set can be empty if there is no movie rated by both u and v. Since the data set is sparse this is likely.

Q9: Can you explain the reason behind mean-centering the raw ratings in the prediction function?

We mean centre the raw ratings to take into consideration the bias with which each user rates a movie. Users tend to rate all movies with a high value or a low value. We should take into consideration the relative rating between users rather than their absolute raw values. For eg: Users A and B might like a movie to the same extent but one might give a rating of 3 and the other 4.

Q10: k-NN collaborative filter using 10-fold cross validation (k=2 to 100)

The average RMSE and average MAE values that we observed for each k are as shown in the table below:

K	Average RMSE	Average MAE
2	1.0530770120109612	0.81176165381982435
4	0.97861907193292175	0.75183679101612122
6	0.95116566966823668	0.72965822928341706
8	0.94029377666341707	0.71952868951438786
10	0.93374035326850358	0.71360530707302106
12	0.92920770678165066	0.70940168606658272
14	0.92590920935945031	0.70716294779545374
16	0.92285258060461572	0.70451843670051206
18	0.92356453671226935	0.70450292283169957
20	0.92191555180623919	0.70261795551329587
22	0.92109466017762431	0.70195339375212518
24	0.91992721854600923	0.70176226788977059

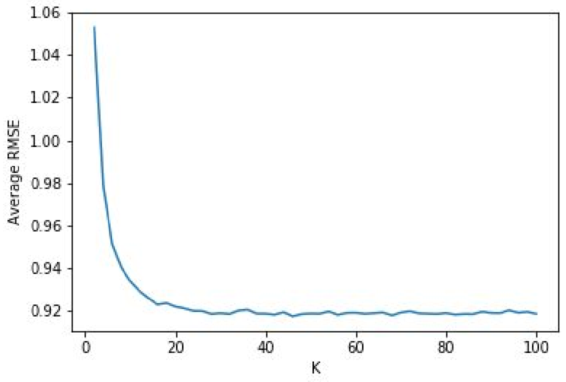


Figure: Average RMSE vs K

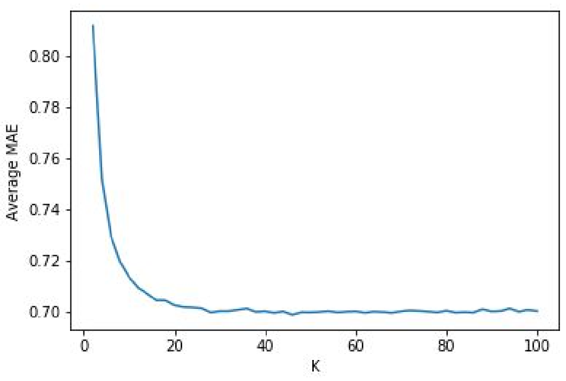


Figure: Average MAE vs K

It can be seen that the average RMSE and MAE values decrease from 0 to 20 and then become stable.

Q11: Minimum ‘k’ corresponding to a steady-state RMSE and MAE value

From the above plot, the observed steady state values of average RMSE and average MAE correspond to 0.92 and 0.70 respectively. The corresponding minimum k value is observed to be 20.

Q12: k-NN collaborative filter using 10-fold cross validation on popular movie trimmed test set (k=2 to 100)

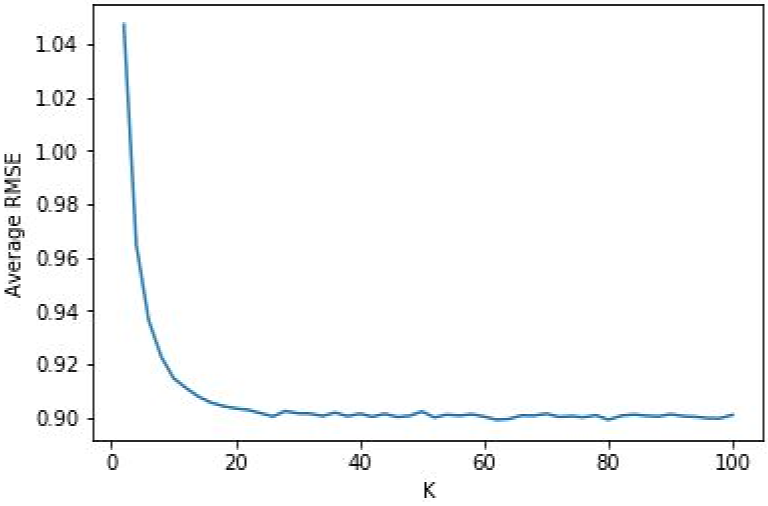


Figure: Average RMSE vs K for popular trimmed test set

Minimum average RMSE: 0.899107659471

Q13: k-NN collaborative filter using 10-fold cross validation on unpopular movie trimmed test set (k=2 to 100)

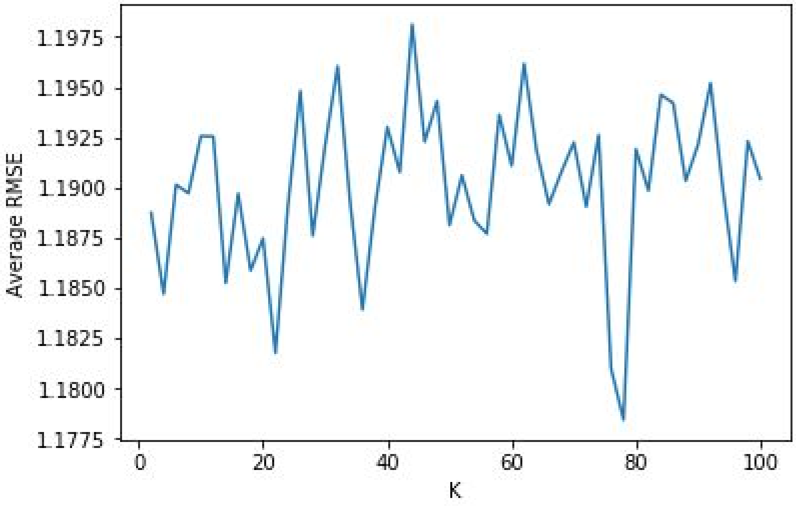


Figure: Average RMSE vs K for unpopular trimmed test set

Minimum average RMSE: 1.1784322993

Q14: k-NN collaborative filter using 10-fold cross validation on high variance movie trimmed test set (k=2 to 100)

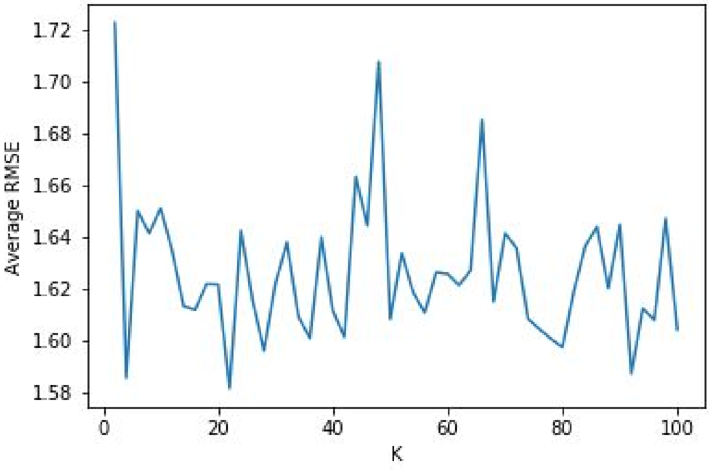
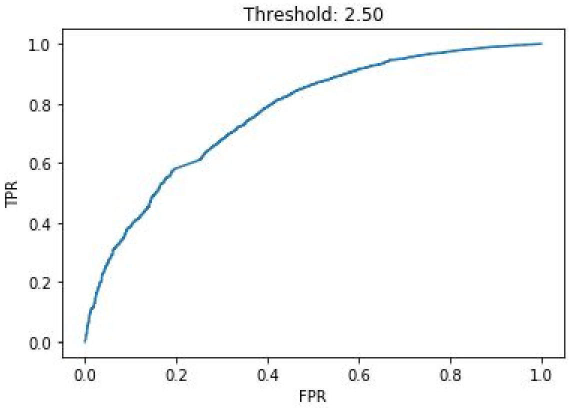


Figure: Average RMSE vs K for high variance trimmed test set

Minimum average RMSE: 1.58121032567

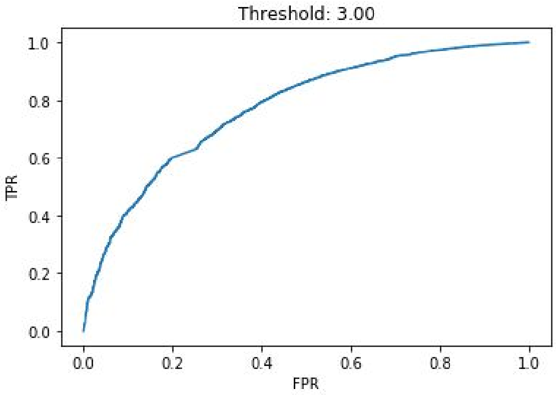
Q15: ROC curves for various threshold values (Q10, k=20)

Threshold: 2.5



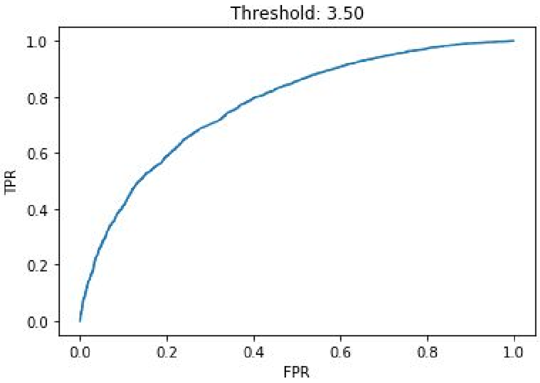
Area under the ROC curve: 0.767938934692

Threshold: 3.0



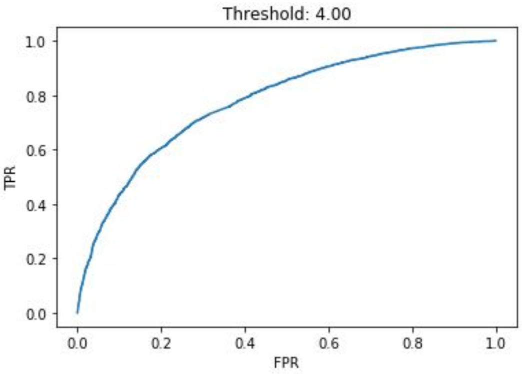
Area under the ROC curve: 0.773866233401

Threshold: 3.5



Area under the ROC curve: 0.774249945241

Threshold: 4.0



Area under the ROC curve: 0.778350703781

Q16: Is the optimization problem given by equation 5 convex? Consider the optimization problem given by equation 5. For U fixed, formulate it as a least-squares problem.

$$\underset{U,V}{\text{minimize}} \quad \sum_{i=1}^{m}\sum_{j=1}^{n} W_{ij}(r_{ij} - (UV^T)_{ij})^2 \qquad (5)$$

The optimization problem given by equation 5 is non-convex. The cost function is non-convex because both U and V are unknowns. To make this optimization problem convex, we can fix U and solve for V and fixed V and solve for U.

By keeping the 'U' fixed, the n rows of V are each treated as a least square model and V is solved. In order to determine the optimal vector, say $\overline{Vj}$, for each row j, the following equation is minimized

$$\sum_{i:(i,j)\in S} W_{ij}\,(r_{ij} - \sum_{s=1}^{k} u_{is}v_{js})^2$$

Thus, we have a least square regression problem in v_{j1},v_{j2}, etc. while u_{i1},u_{i2} are considered constants and v_{j1} and v_{j2} are the optimization variables.

Q17: NNMF collaborative filter using 10-fold cross validation (k=2 to 50)

The average RMSE and average MAE values that we observed for each k are as shown in the table below:

K	Average RMSE	Average MAE
2	1.1773728647584902,	0.99555624656187702
4	1.070621469751889	0.88035044499761883
6	1.0122901999442271	0.81490225879036182
8	0.97649386684495809	0.77308803629631728

10	0.9541799035771108	0.74630777450233221
12	0.94739732321110393	0.73420454606756058,
14	0.94154377277041534	0.72594671014455381
16	0.93786163127667899	0.71910415915154902
18	0.93700221090647362	0.71398096819545265,
20	0.93984121327810399	0.71404132889846239
22	0.94357776718515873	0.71480710545644666,
24	0.9453247409027139	0.71474734087928427
26	0.94738114250275629	0.71491932358377985
28	0.95291588701219732,	0.7183450794638786

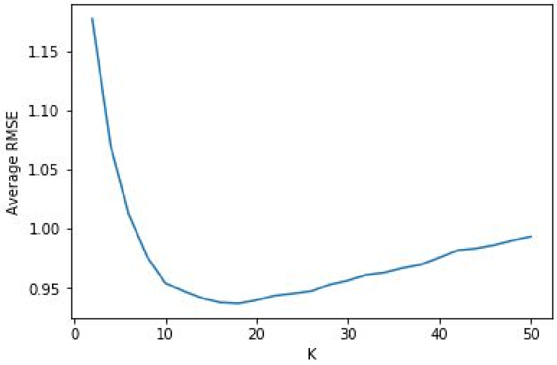


Figure: Average RMSE vs K

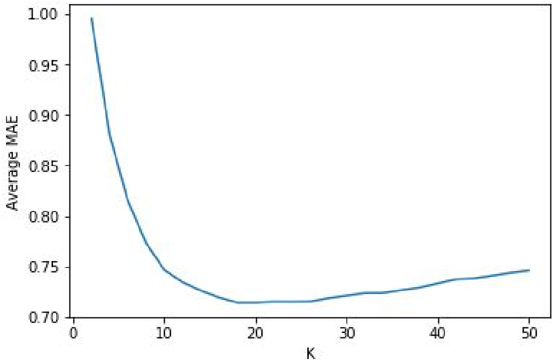


Figure: Average MAE vs K

Q18: Optimal number of latent factors

The optimal number of latent factors is 20 which is equal to the number of movie genres.

Q19: NNMF collaborative filter using 10-fold cross validation on popular movie trimmed test set (k=2 to 50)

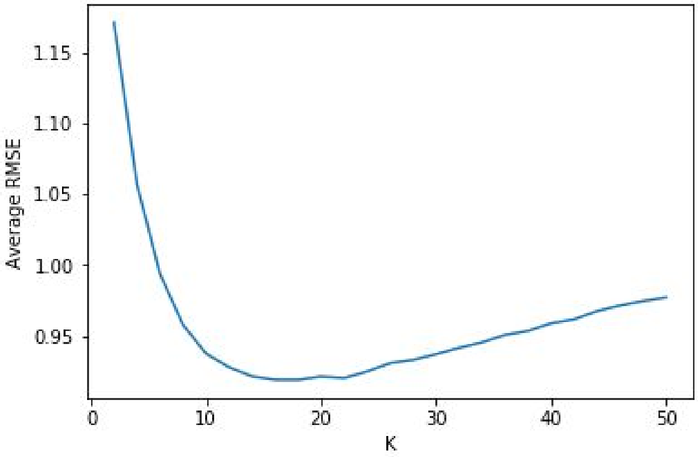


Figure: Average RMSE vs K

Minimum average RMSE: 0.919054194517

Q20: NNMF collaborative filter using 10-fold cross validation on unpopular movie trimmed test set (k=2 to 50)

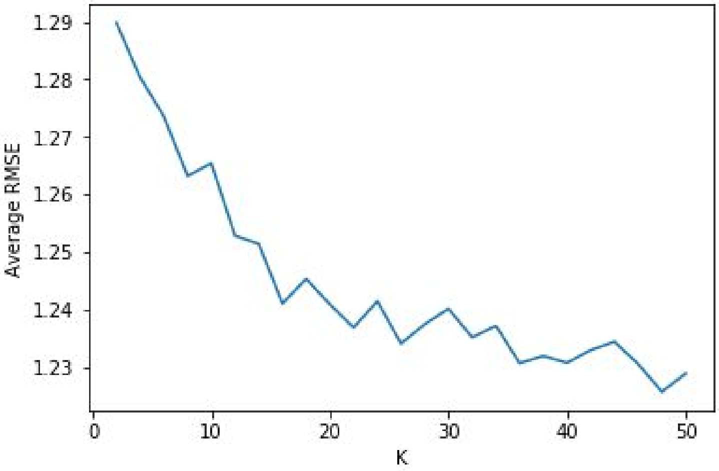


Figure: Average RMSE vs K for unpopular trimmed test set

Minimum average RMSE: 1.22568546557

Q21: NNMF collaborative filter using 10-fold cross validation on high variance movie trimmed test set (k=2 to 50)

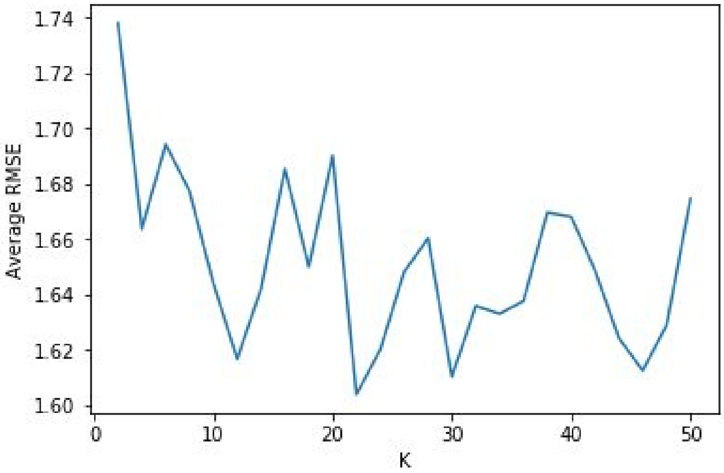
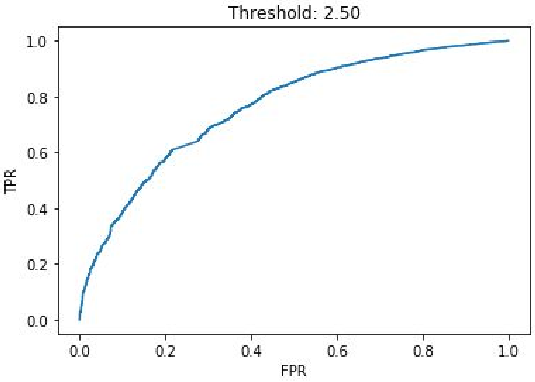


Figure: Average RMSE vs K for high variance trimmed test set

Minimum average RMSE: 1.60408033118

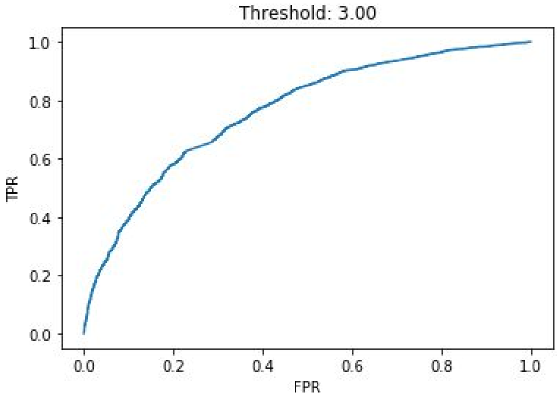
Q22: ROC curves for the NNMF collaborative filter for different thresholds.

Threshold: 2.5



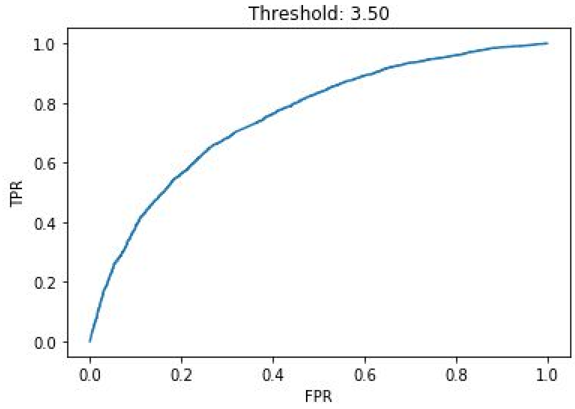
Area under the ROC curve: 0.761871525875

Threshold: 3.0



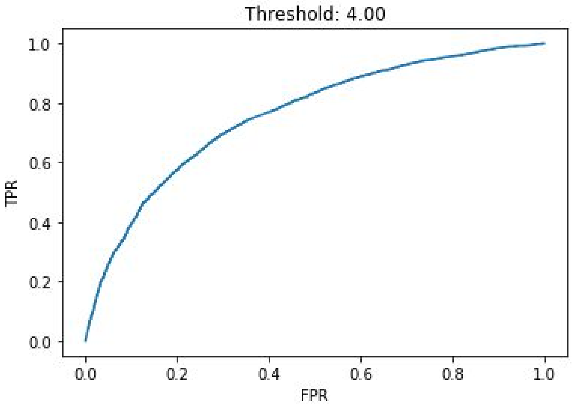
Area under the ROC curve: 0.763712328655

Threshold: 3.5



Area under the ROC curve: 0.754819036692

Threshold: 4.0



Area under the ROC curve: 0.7579035658

Q23: Latent factors and movie genres

The genres of the top 10 movies

Column 1:

Drama|Mystery|Romance
Action|Comedy|Crime|Fantasy
Comedy|Documentary
Drama|War
Children|Comedy
Action|Adventure|Sci-Fi|War|IMAX
Thriller
Comedy|Western
Action|Adventure|Sci-Fi|IMAX
Drama

Column 2:

Comedy|Documentary
Adventure|Animation
Musical
Comedy|Romance
Drama|Fantasy|Horror

Comedy
Drama|Sci-Fi
Drama
Adventure|Drama|Sci-Fi
Drama|Romance

Column 3:

Adventure|Drama|Fantasy|Romance
Action|Adventure|Drama|Thriller
Action|War
Action|Crime|Thriller
Comedy
Comedy|Drama|Romance
Comedy
Comedy|Mystery|Thriller
Drama
Comedy

Column 4:

Drama
Comedy|Crime
Comedy|Drama
Documentary|Drama
Action
Drama|Thriller
Adventure|Children
Drama|Romance
Comedy
Drama

Column 5:

Comedy|Crime|Mystery|Thriller
Action|Comedy
Documentary
Action|Adventure|Sci-Fi|Thriller
Comedy|Drama

Drama|Romance
Children|Comedy|Musical|Romance
Comedy
Documentary
Comedy|Romance

It can be observed that the top 10 movies belong to a small collection of genres like Drama, Comedy, Action, Sci-fi and Crime.

The latent factors are hidden or inherent features in a users taste that influences the recommended item. In this case, the movie genres are these hidden features that drive the users to rate movies of a particular genre, towards which they are biased, in a similar manner. Thus 9066 ratings could be replaced by 20 ratings, which is the number of genres, for a user, as all movies of a genre are rated similarly by a user.

Q24: MF with bias collaborative filter using 10-fold cross validation (k=2 to 50)

The average RMSE and average MAE values that we observed for each k are as shown in the table below:

K	Average RMSE	Average MAE
2	1.1768397994899087	0.99538487880511828
4	1.0709124907298202	0.88090882516878943
6	1.0120777865229202	0.81463715103680401
8	0.97557452903278752	0.77167518121140366
10	0.9580975744861997	0.74894642296312708
12	0.94694214230363483	0.73486480219087535
14	0.94175563867956469	0.72582376806639692
16	0.93994399504052883	0.72043420428687432
18	0.93941601668091901	0.71654680226771994
20	0.94270830119290205	0.71570376006305092

22	0.94178398395112384	0.71414258311387047
24	0.94676834273666066	0.71552058345698732
26	0.94625443385398833	0.71454621884243363
28	0.95086752927117835	0.71570376006305092

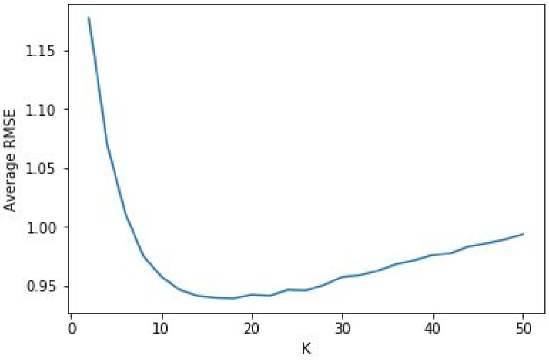


Figure: Average RMSE vs K

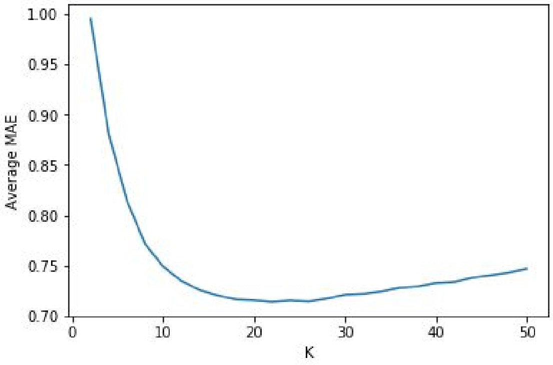


Figure: Average MAE vs K

Q25: Optimal number of latent factors

The minimum values of average RMSE and average MAE are 0.941783983951 and 0.714142583114 and the corresponding optimal k is 20

Q26: MF with bias collaborative filter using 10-fold cross validation on popular trimmed test set (k=2 to 50)

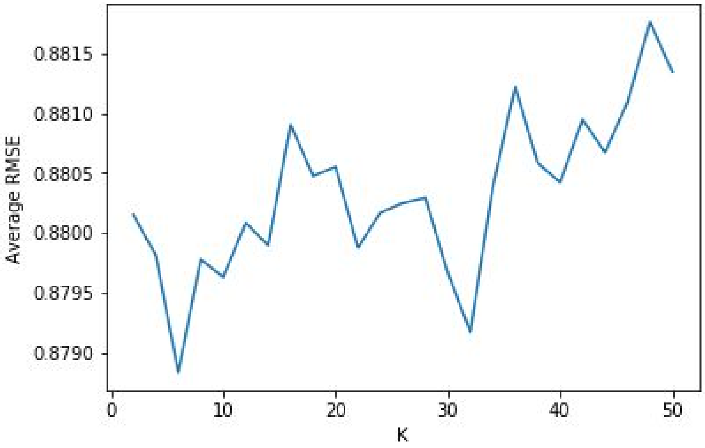


Figure: Average RMSE vs K for popular trimmed test set

Minimum average RMSE: 0.878833173986

Q27: MF with bias collaborative filter using 10-fold cross validation on unpopular trimmed test set (k=2 to 50)

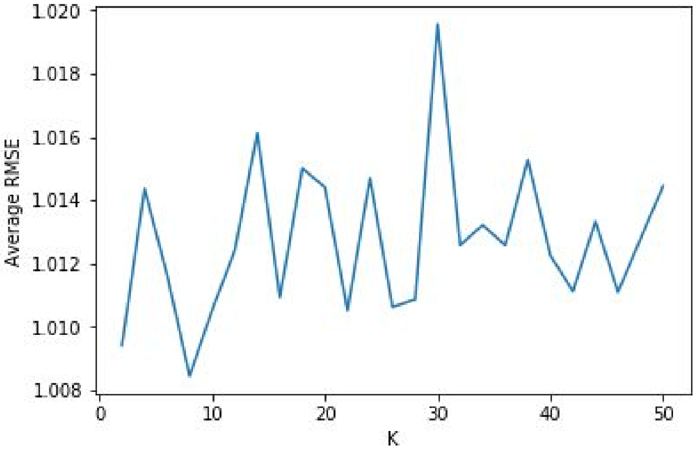


Figure: Average RMSE vs K for unpopular trimmed test set

Minimum average RMSE: 1.00845151548

Q28: MF with bias collaborative filter using 10-fold cross validation on high variance trimmed test set (k=2 to 50)

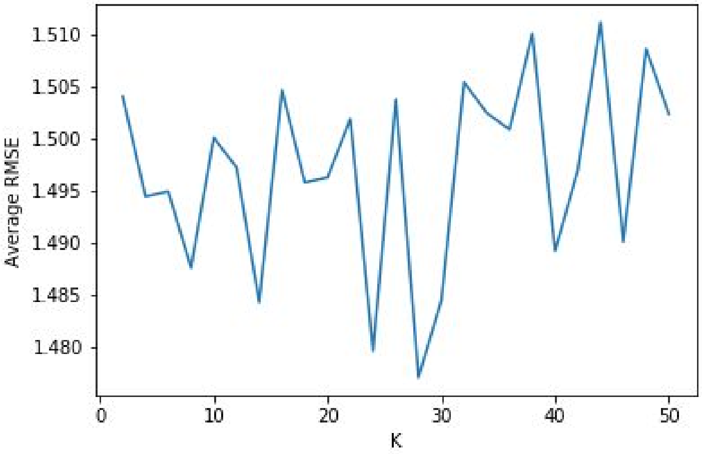
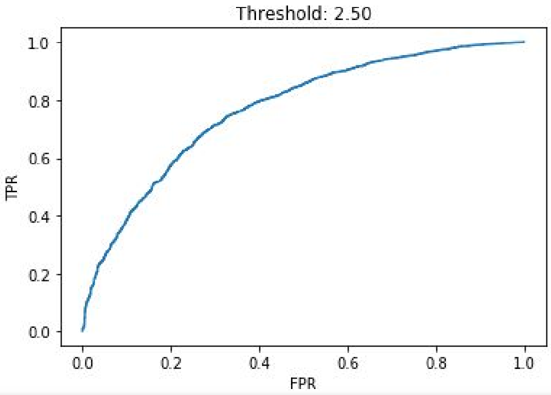


Figure: Average RMSE vs K for high variance trimmed test set

Minimum average RMSE: 1.4771089218

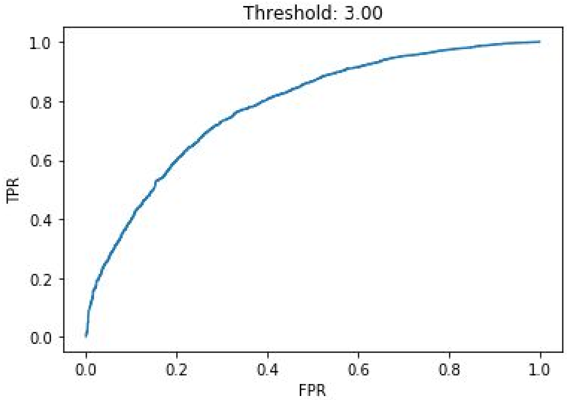
Q29: ROC curves for MF with bias collaborative filter for various thresholds.

Threshold: 2.5



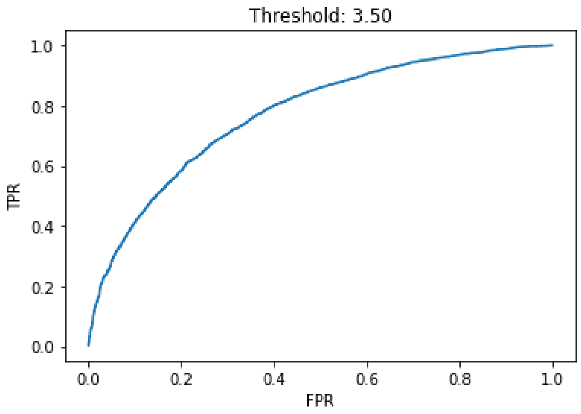
Area under the ROC curve: 0.769530226079

Threshold: 3.0



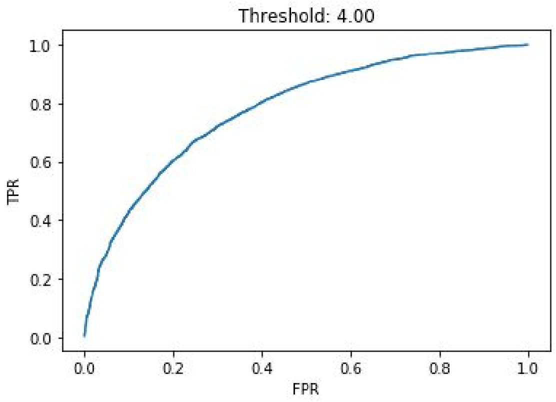
Area under the ROC curve: 0.780293202469

Threshold: 3.5



Area under the ROC curve: 0.773677553877

Threshold: 4.0



Area under the ROC curve: 0.78089515578

Q30: Naive collaborative filter using 10-fold cross validation

Average RMSE: 0.9553960518829795

Q31: Naive collaborative filter using 10-fold cross validation on popular movie trimmed test set

Average RMSE: 0.952119547547865

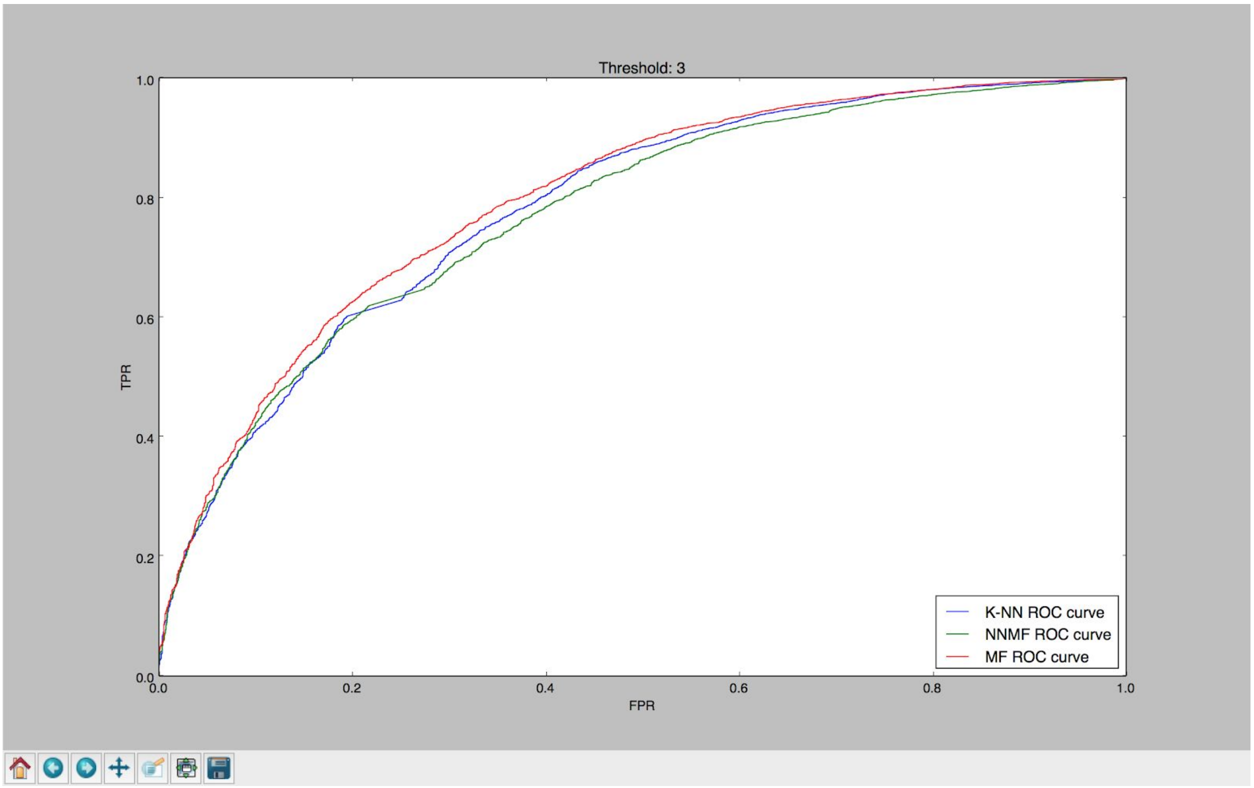
Q32: Naive collaborative filter using 10-fold cross validation on unpopular movie trimmed test set

Average RMSE: 1.0101017740731004

Q33: Naive collaborative filter using 10-fold cross validation on high variance movie trimmed test set

Average RMSE: 1.5094979647532187

Q34: ROC curves (threshold = 3) for the k-NN, NNMF, and MF with bias based collaborative filters in the same figure

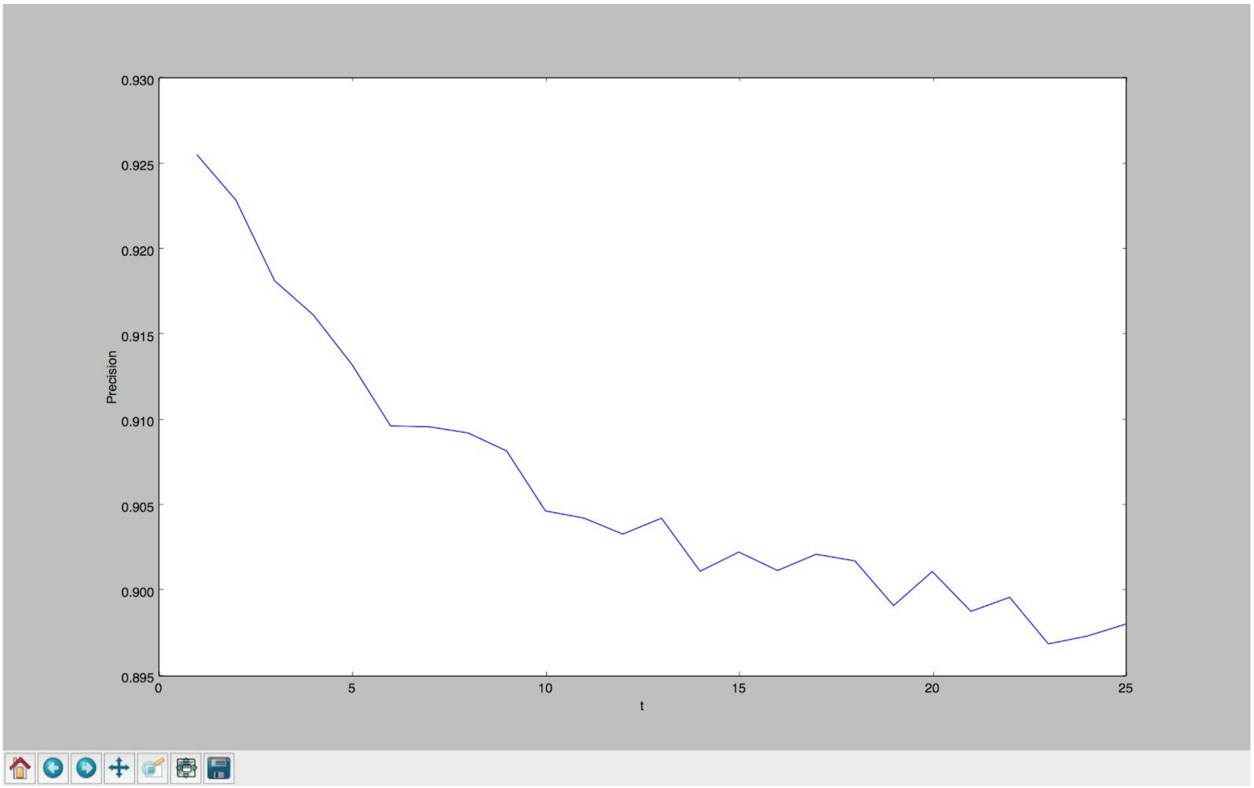


It can be seen from the graph above that the best performance can be achieved by using MF with bias based collaborative filter since it has the highest value for area under the curve.

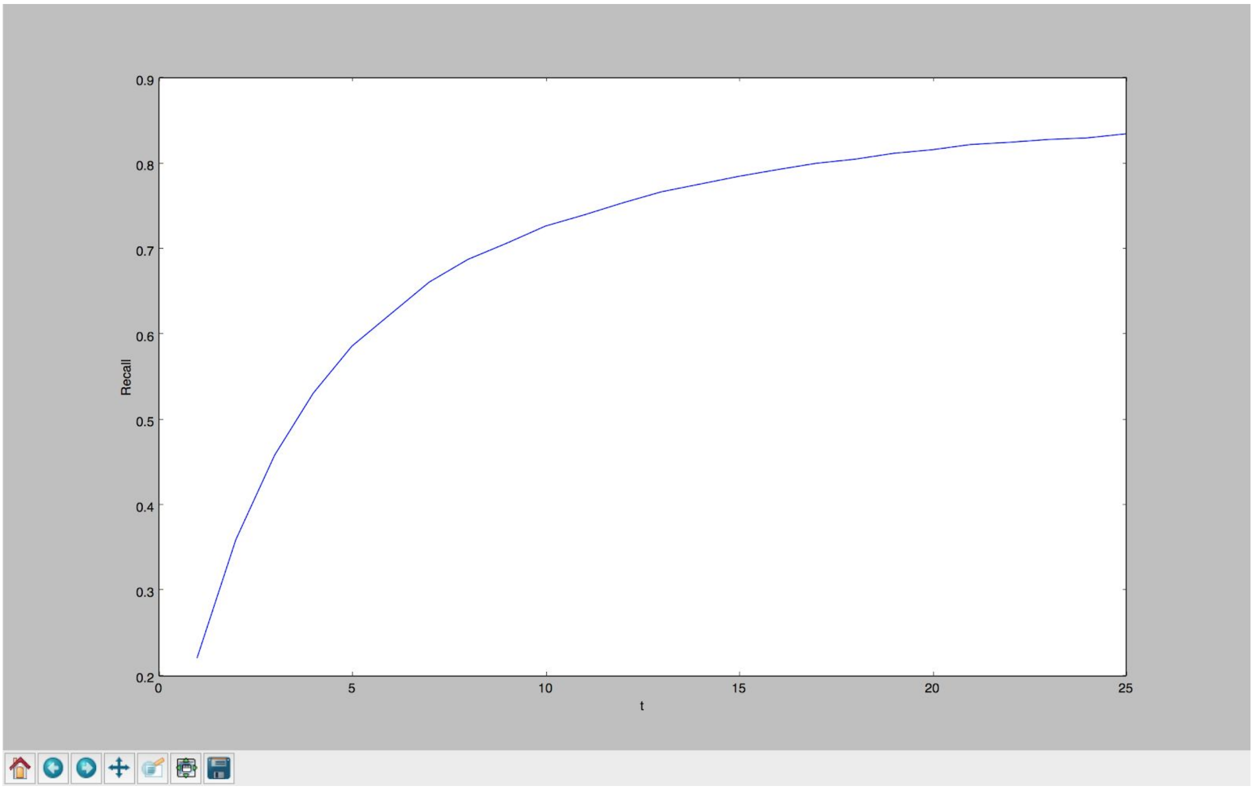
Q35: Explain the meaning of precision and recall in your own words

Precision is the number of movies retrieved that are relevant and Recall is the number of relevant movies that are retrieved. Precision and Recall are both extremely useful in understanding what set of documents or information was presented and how many of those documents are actually useful to the question being asked. Precision and Recall are inversely proportional to each other.

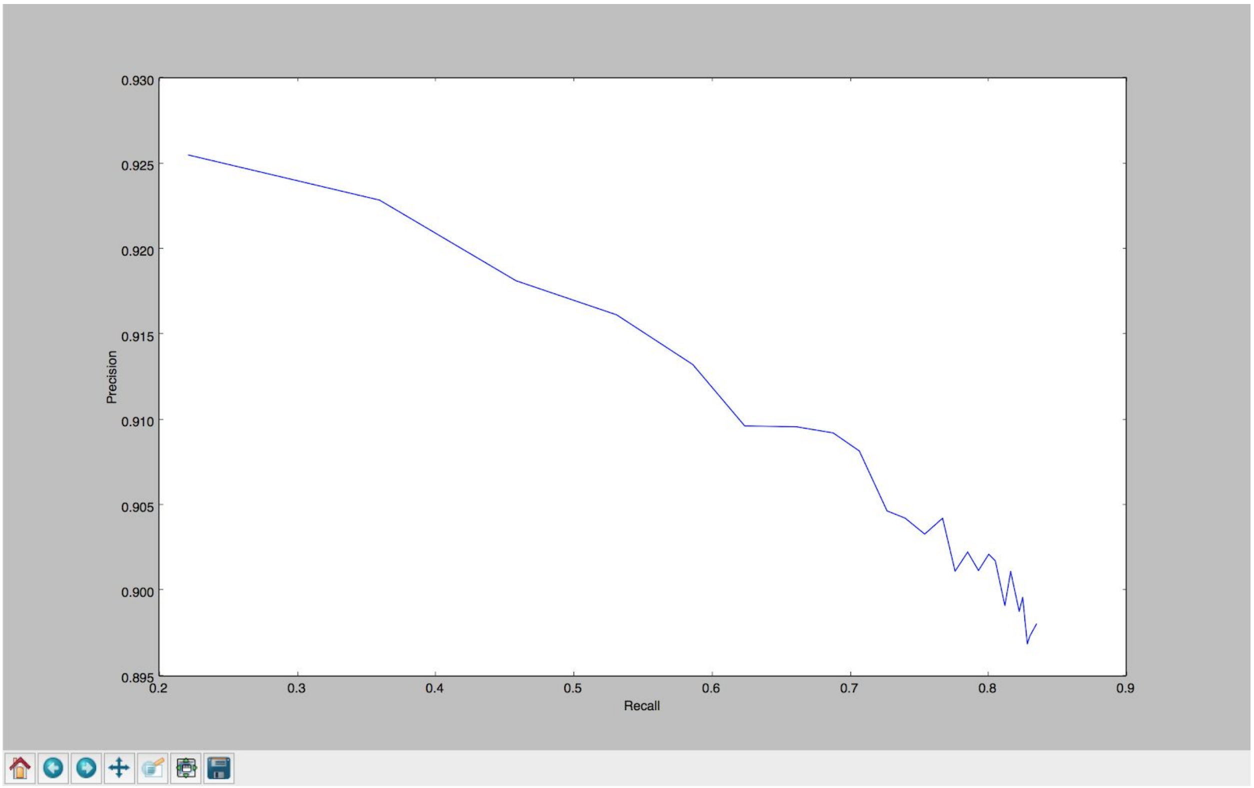
Q36: Plot average precision (Y-axis) against t (X-axis) for the ranking obtained using k-NN collaborative filter predictions. Also, plot the average recall (Y-axis) against t (X-axis) and average precision (Y-axis) against average recall (X-axis)



As more movies are recommended the precision keeps dropping.

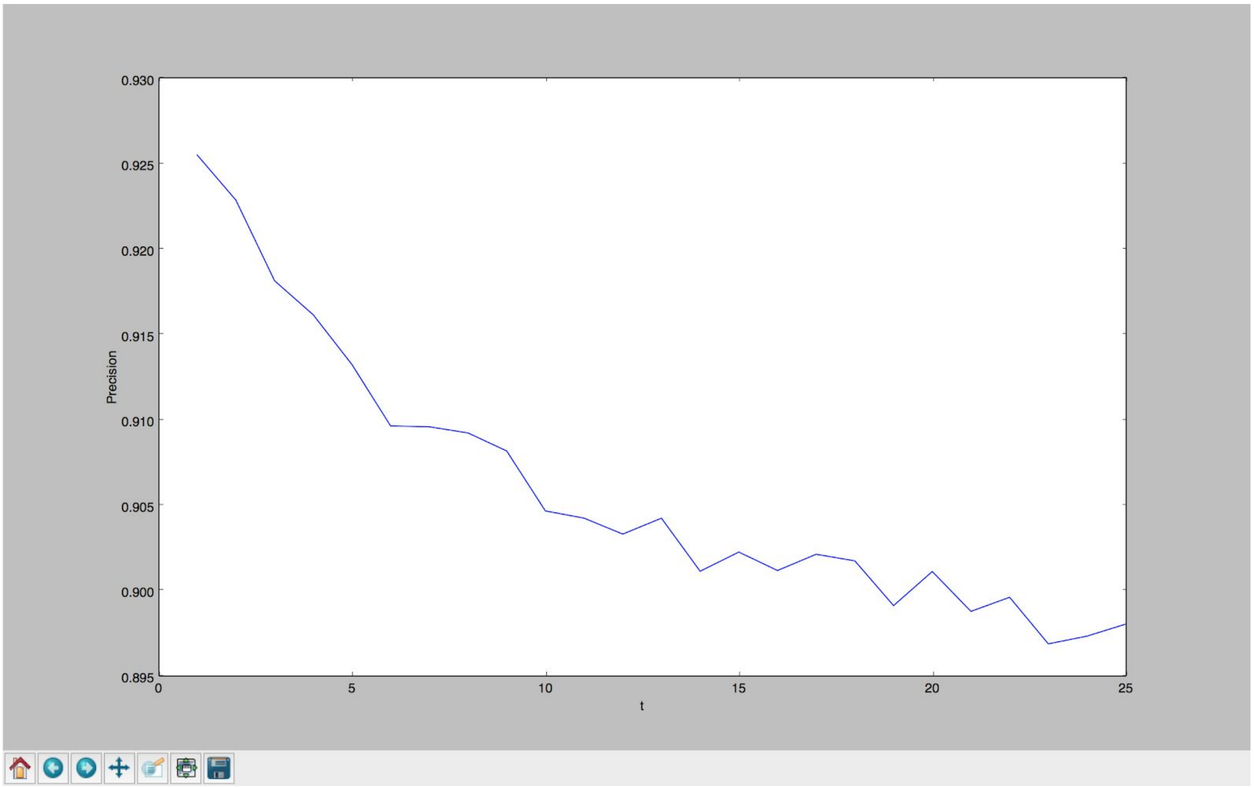


As more movies are recommended the recall keeps on increasing.

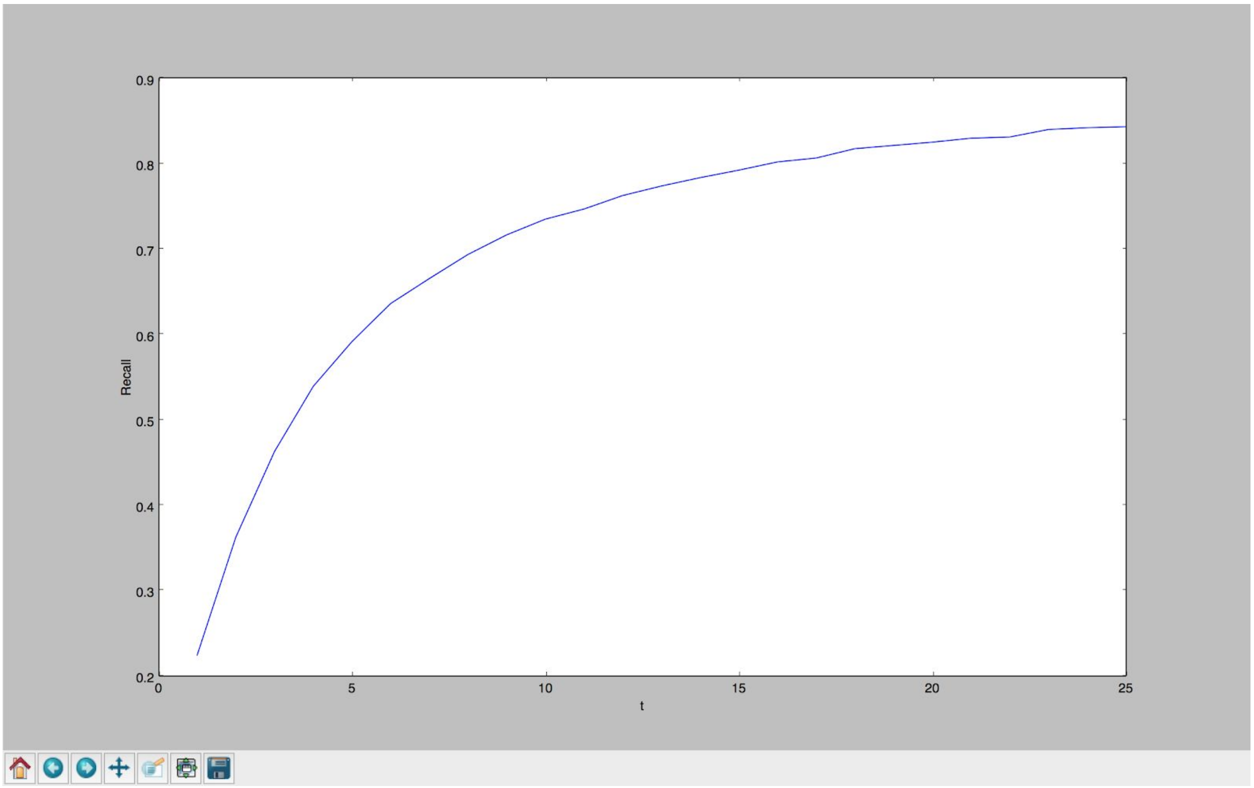


As recall increases, values of precision decreases.

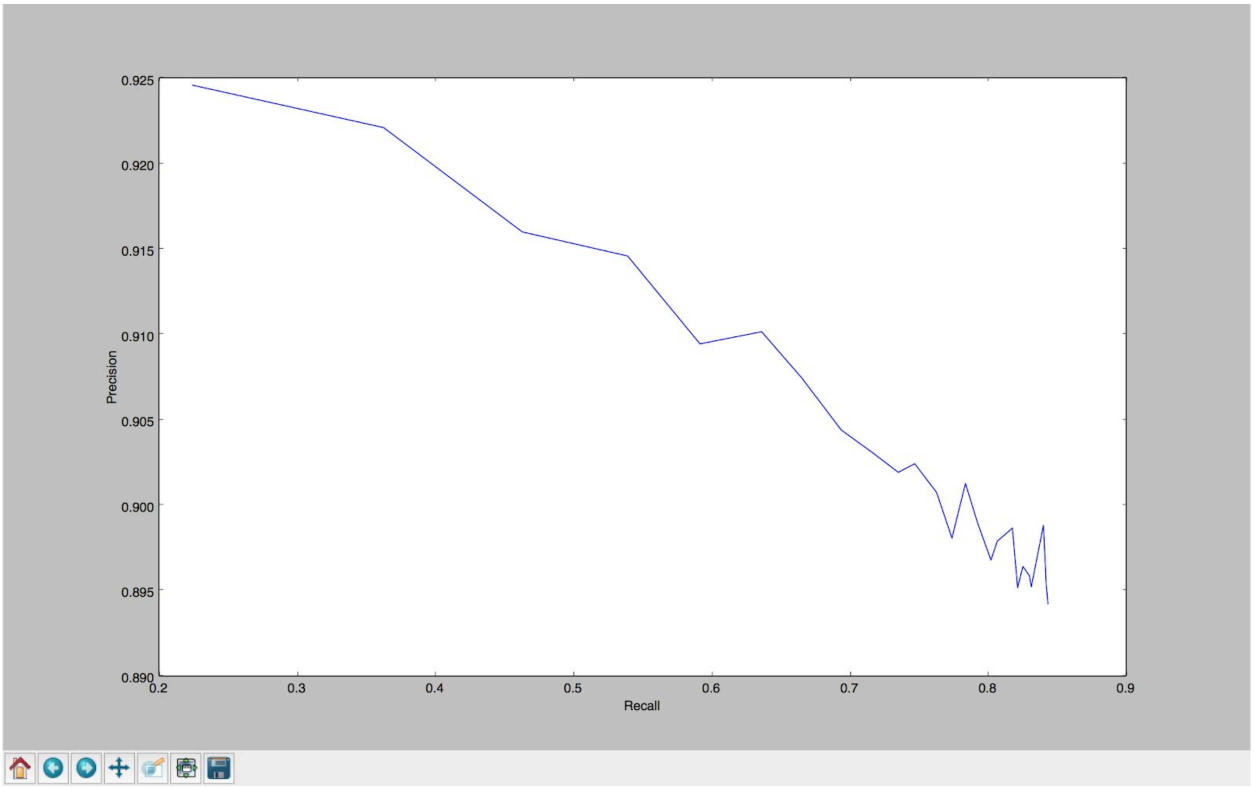
Q37: Plot average precision (Y-axis) against t (X-axis) for the ranking obtained using NNMF-based collaborative filter prediction. Also, plot the average recall (Y-axis) against t (X-axis) and average precision (Y-axis) against average recall (X-axis)



As more movies are recommended the precision keeps dropping.

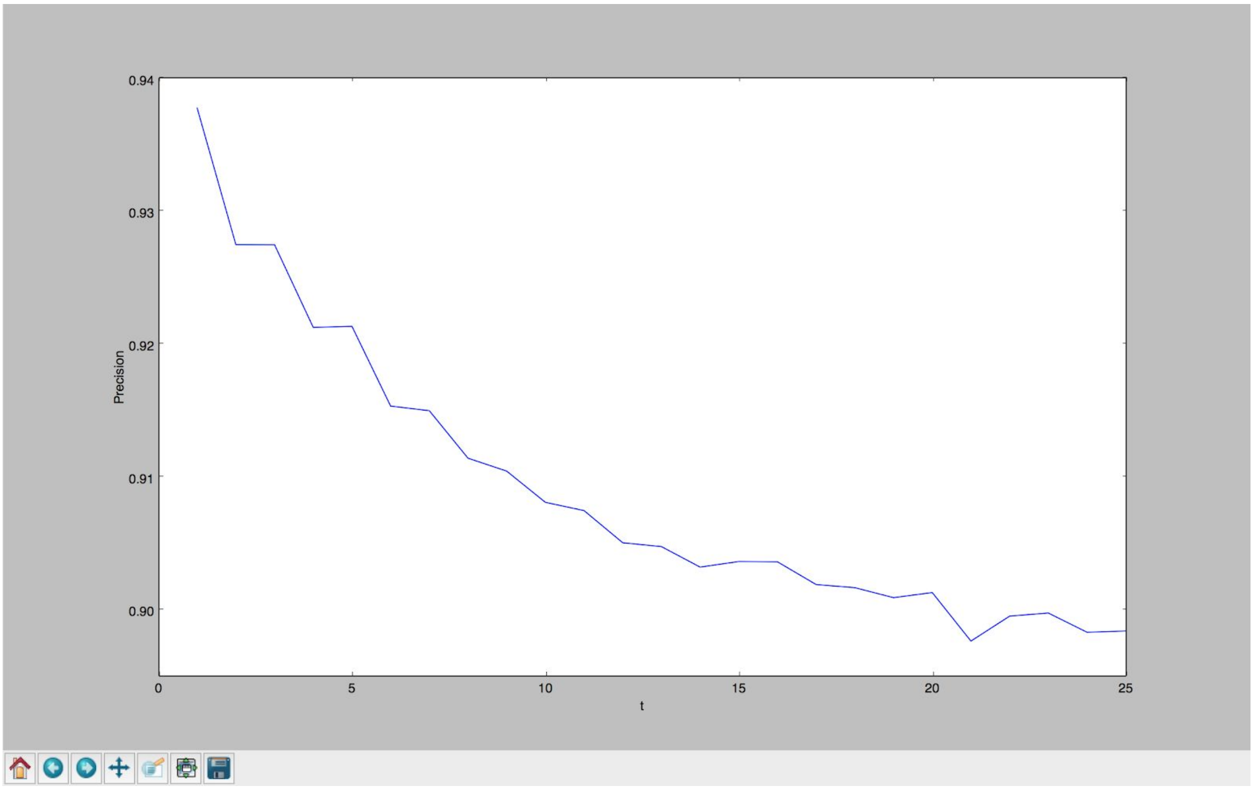


As more movies are recommended the recall keeps on increasing.

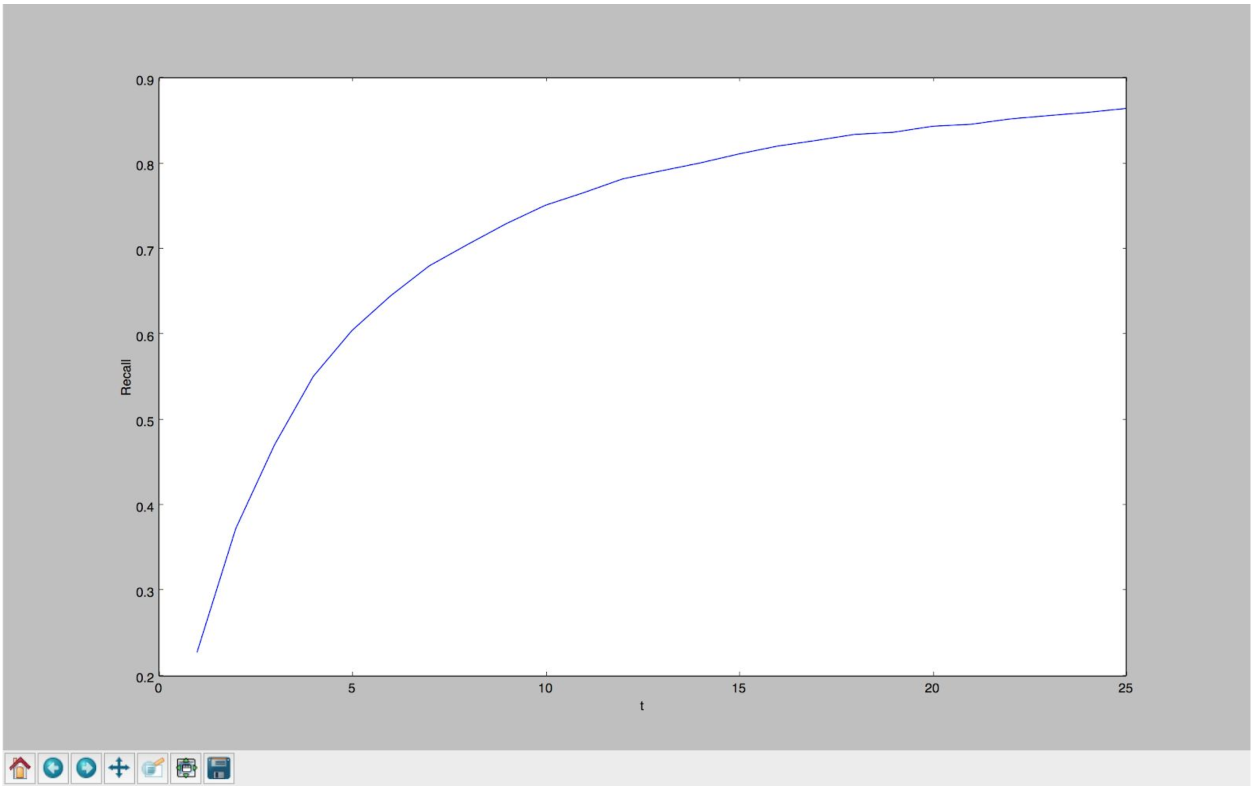


As recall increases, values of precision decreases.

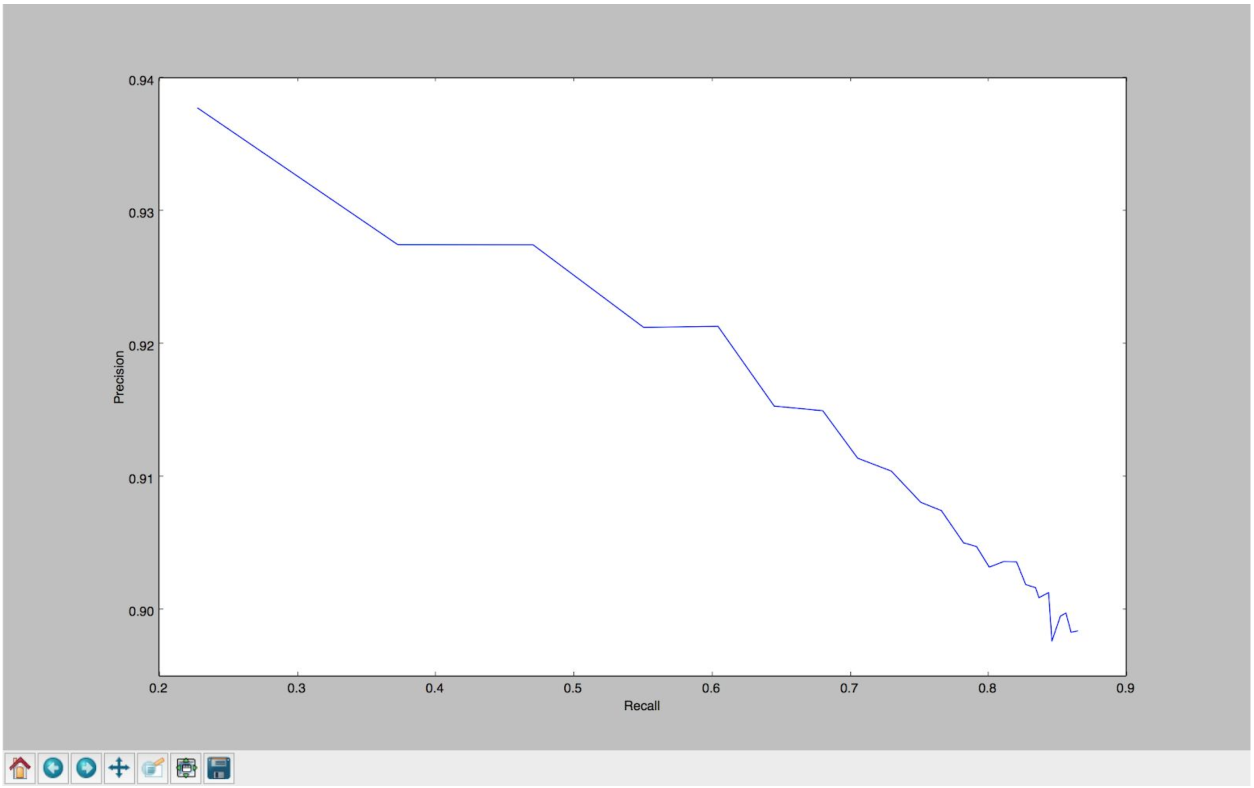
Q38: Plot average precision (Y-axis) against t (X-axis) for the ranking obtained using MF with bias-based collaborative filter prediction. Also, plot the average recall (Y-axis) against t (X-axis) and average precision (Y-axis) against average recall (X-axis)



As more movies are recommended the precision keeps dropping.

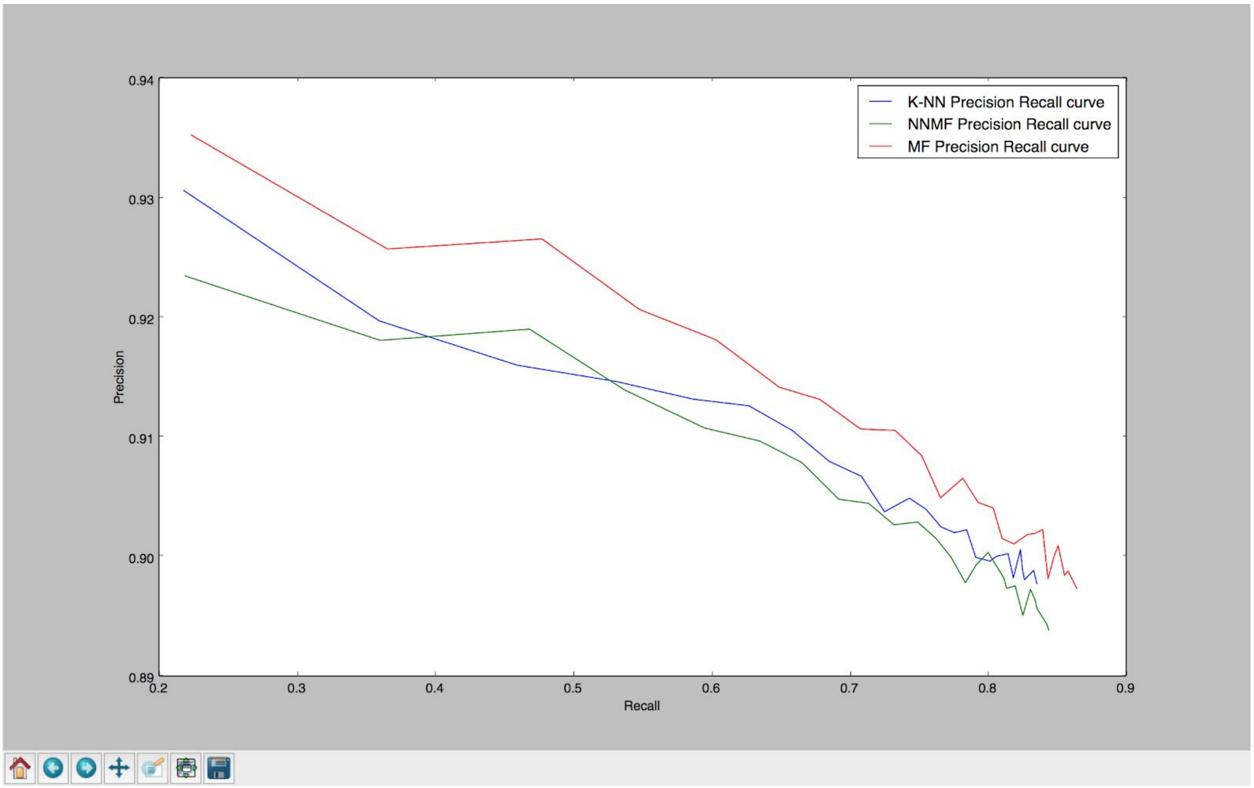


As more movies are recommended the recall keeps on increasing.



As recall increases, values of precision decreases.

Q39: Plot the precision-recall curve obtained using k-NN, NNMF, and MF with bias predictions in the same figure.



It can be seen from the graph above that the best performance can be achieved by using MF with bias based collaborative filter as it provides recommendations which are the most relevant.