EE 219: Large-Scale Data Mining: Models and Algorithms

Project 3: Collaborative Filtering

Akshay Sharma (504946035)

Anoosha Sagar (605028604)

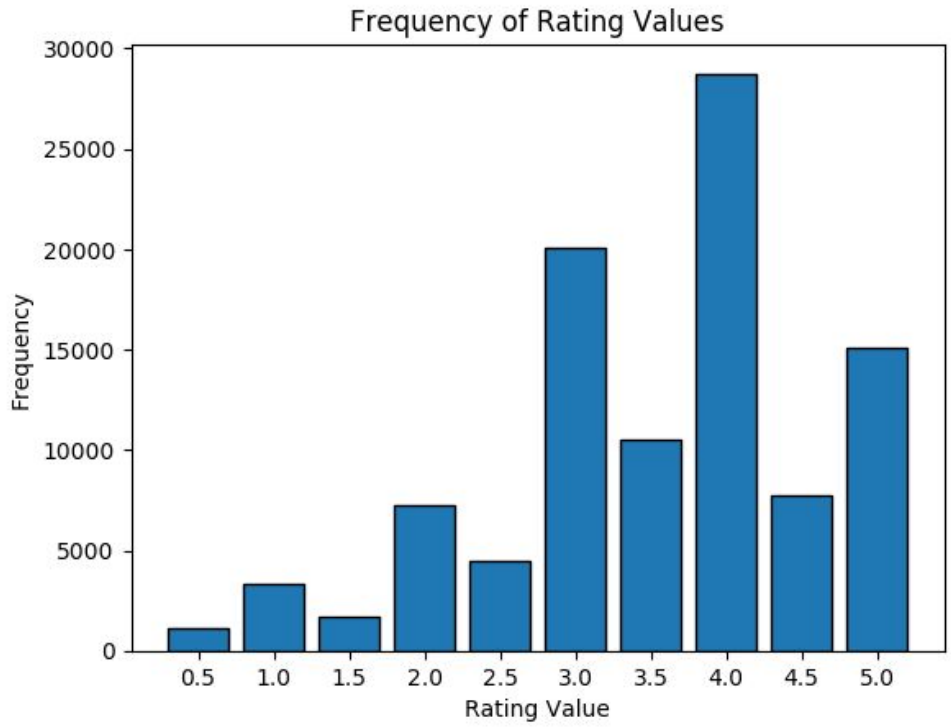
Nikhil Thakur(804946345)

Rahul Dhavalikar (205024839)

Q1) Compute the sparsity of the movie rating dataset.

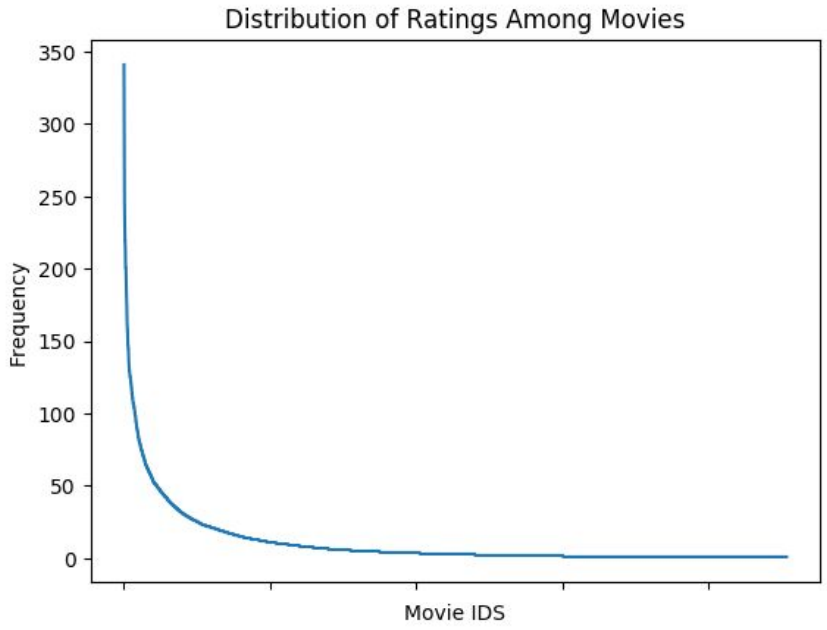
Total Number of available ratings	100004
Total Number of possible ratings	9125 x 671 = 6122875
Sparsity	0.9836671498274911

Q2) Plot a histogram showing the frequency of the rating values. Briefly comment on the shape of the histogram

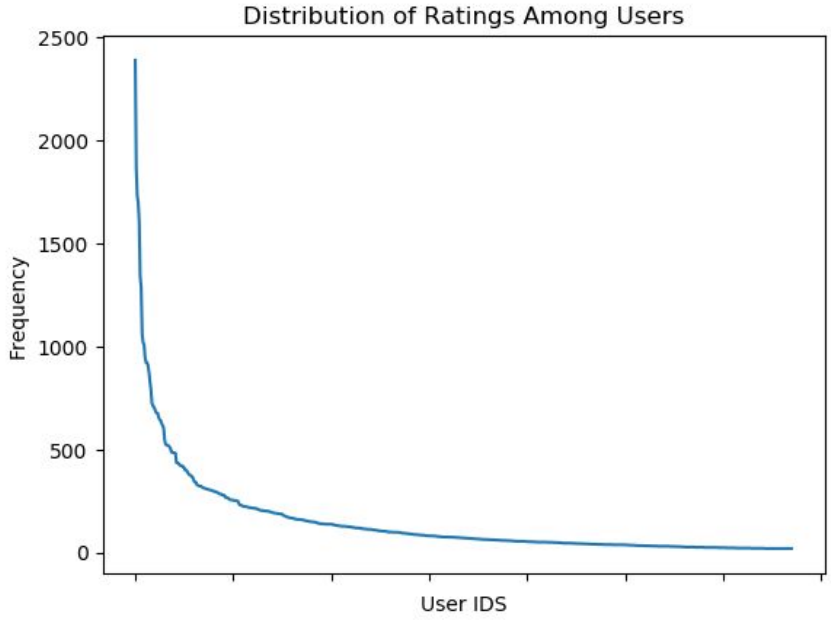


The histogram is very unevenly distributed and most of the ratings are either 3 or greater than 3.

Q3) Plot the distribution of ratings among movies.



Q4) Plot the distribution of ratings among users.

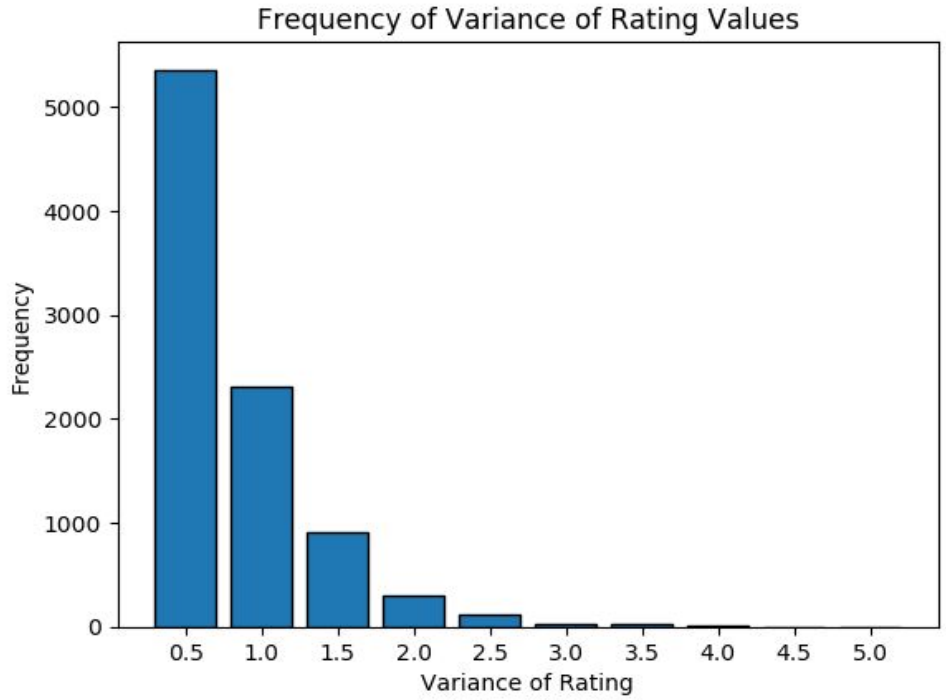


Q5) Explain the salient features of the distribution found in question 3 and their implications for the recommendation process.

We see that only a few of the movies have a lot of ratings. Majority movies on the other hand have only a few ratings.

Consider a scenario where the movie that received the most rating counts was not highly rated at all. As a result, if we were to use recommendations based on rating counts, we would definitely make mistakes here by recommending a movie with poor ratings. So, we need to take care of this while building our system.

Q6) Compute the variance of the rating values received by each movie and plot histogram. Briefly comment on the shape of the histogram



The histogram is unevenly distributed and the count of the ratings decrease as we go to higher variance.

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

$$\mu_u = \frac{\sum_{k \epsilon I_u} r_{uk}}{|I_u|}$$

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

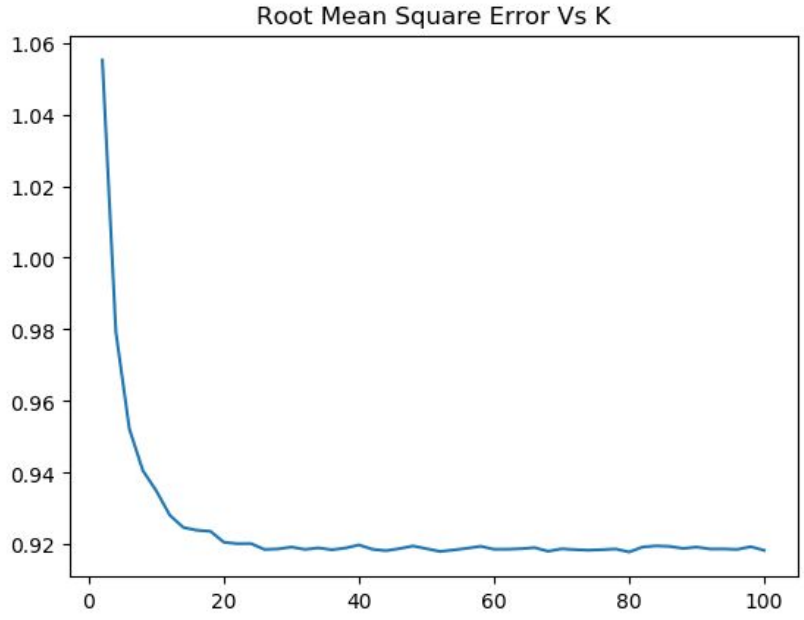
$I_u \cap I_v$ represents the set of items rated by both User 'u' and 'v'

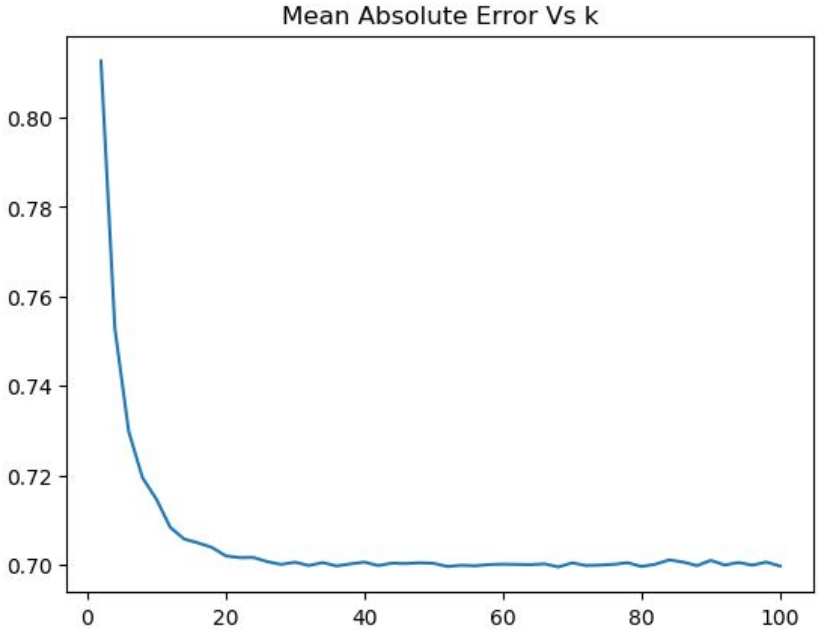
It is possible to $I_u \cap I_v = \varnothing$ to be empty because rating matrices are generally sparse

Q9) Can you explain the reason behind mean-centering the raw ratings (rvj − μv) in the prediction function?

Mean centering is mainly done to reduce the bias in the ratings provided by the users. The problem with Pearson coefficient method is that different users may rate items on different scale. One user may rate all items highly while other users whereas other users may rate all items negatively. This may not give accurate results and introduce and error.

Q10) Design a k-NN collaborative filter to predict the ratings of the movies in the MovieLens dataset and evaluate its performance using 10-fold cross validation. Plot average RMSE (Y-axis) against k (X-axis) and average MAE (Y-axis) against k (X-axis)

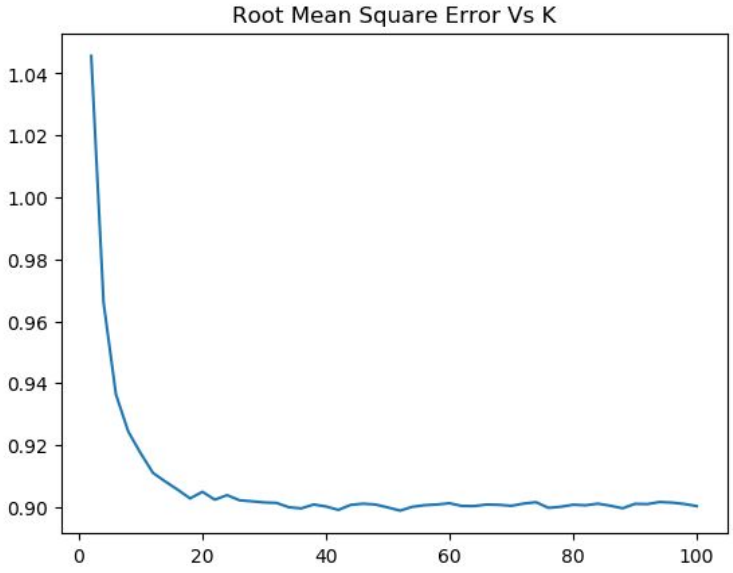


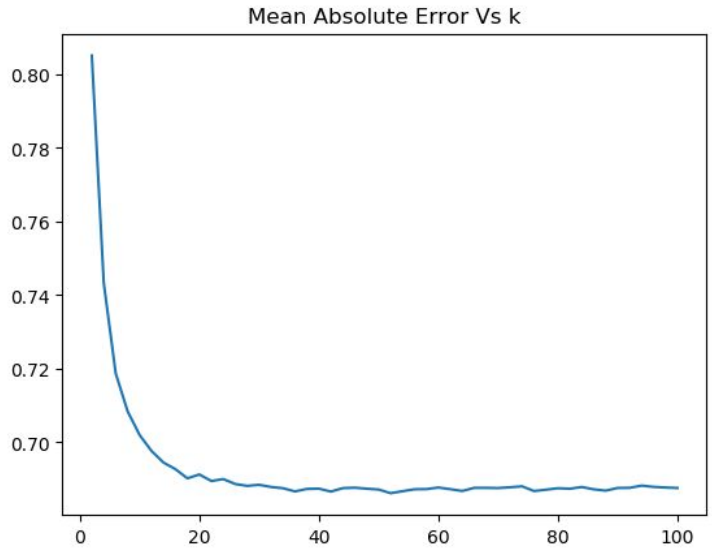


Q11) Use the plot from question 10, to find a 'minimum k'. 'minimum k' corresponds to the k value for which average RMSE and average MAE converges to a steady-state value. Please report the steady state values of average RMSE and average MAE

Minimum k (Steady-state value)	14
Average RMSE at steady state	0.924729197949
Average MAE at steady state	0.705773922421

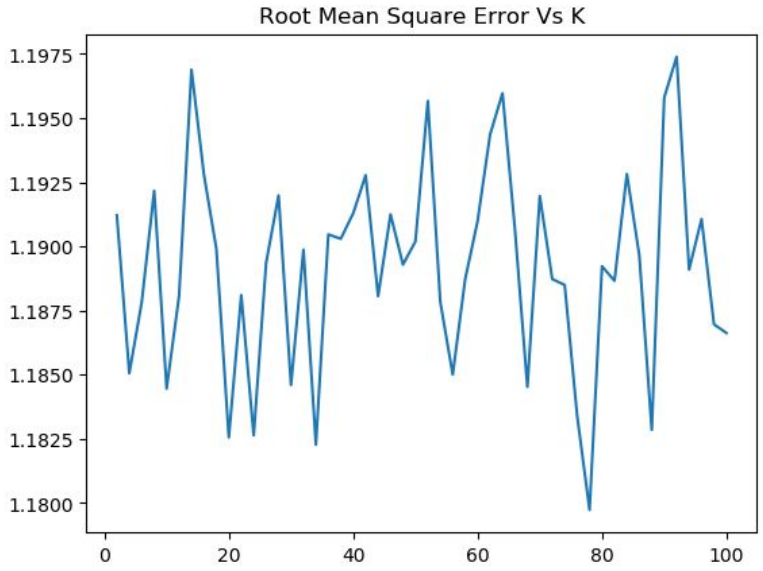
Q12) Design a k-NN collaborative filter to predict the ratings of the movies in the popular movie trimmed test set and evaluate it's performance using 10-fold cross validation.Plot average RMSE (Y-axis) against k (X-axis). Also, report the minimum average RMSE

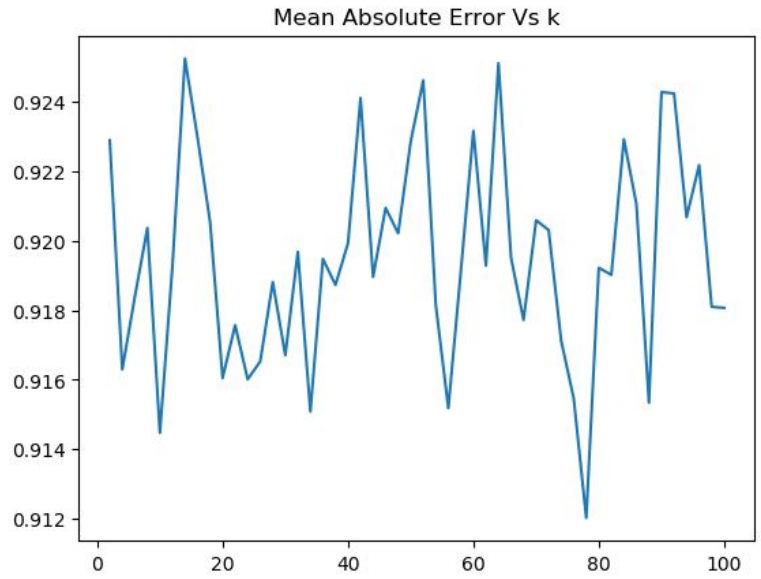




Minimum Average RMSE	0.8988201747148802

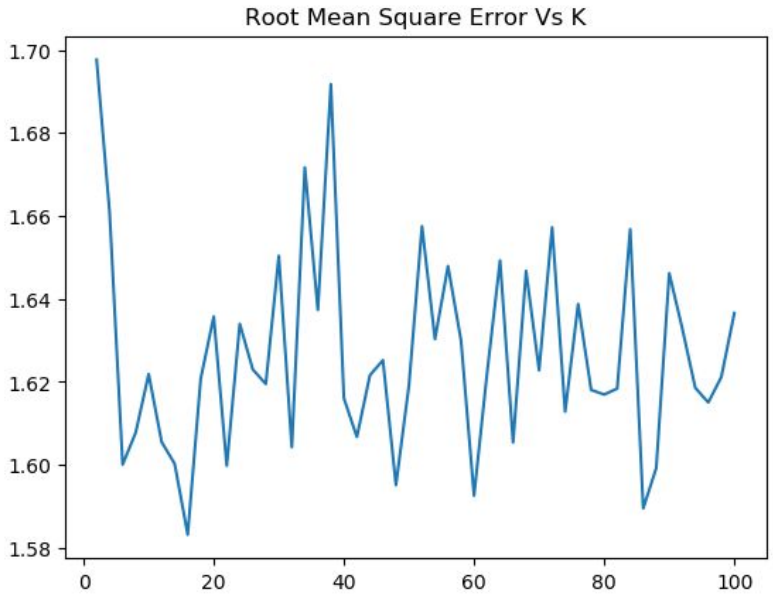
Q13) Design a k-NN collaborative filter to predict the ratings of the movies in the unpopular movie trimmed test set and evaluate it's performance using 10-fold cross validation. Plot average RMSE (Y-axis) against k (X-axis). Also, report the minimum average RMSE

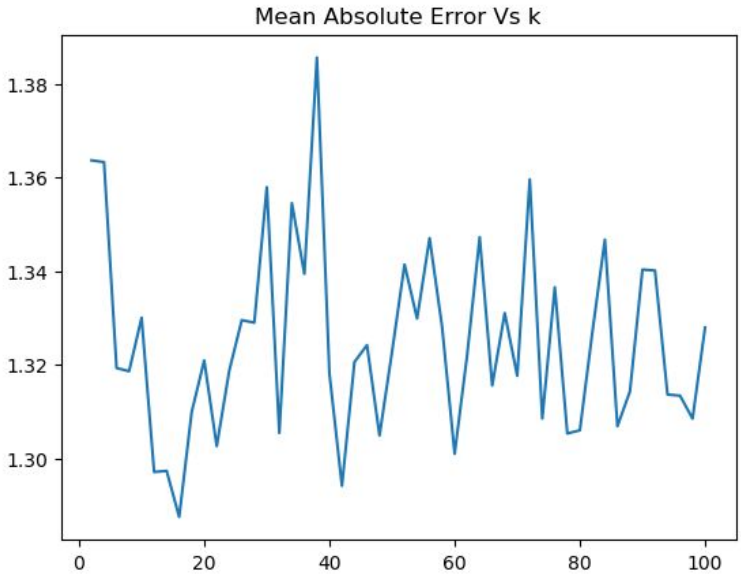




Minimum Average RMSE	1.1797357001037052

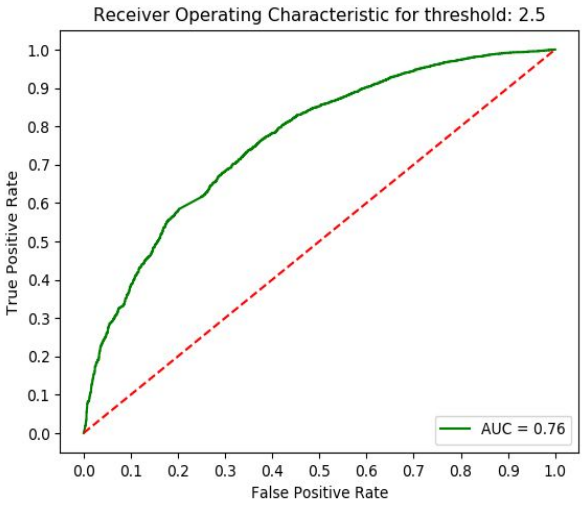
Q14) Design a k-NN collaborative filter to predict the ratings of the movies in the high variance movie trimmed test set and evaluate it's performance using 10-fold cross validation. Plot average RMSE (Y-axis) against k (X-axis). Also, report the minimum average RMSE

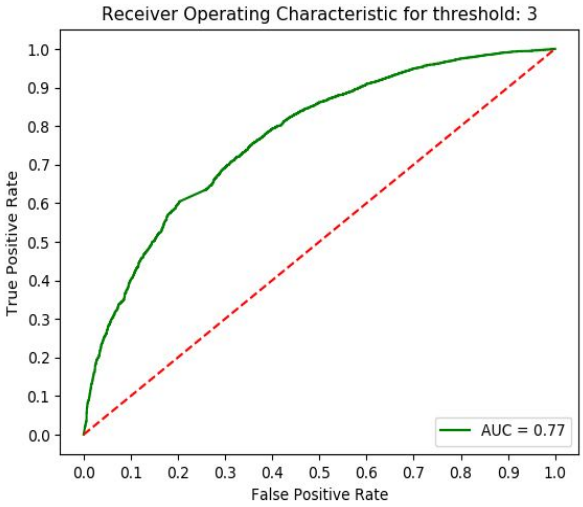


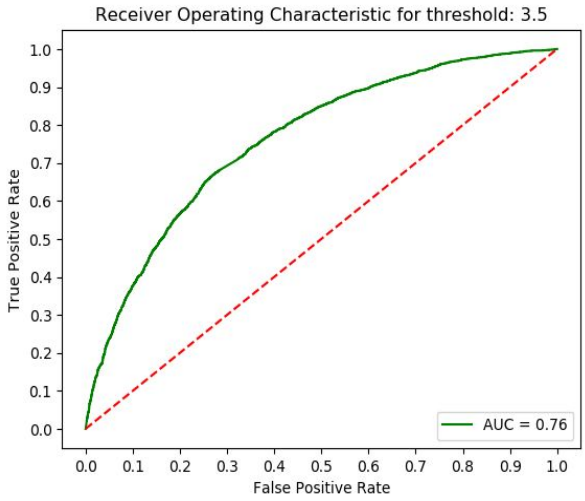


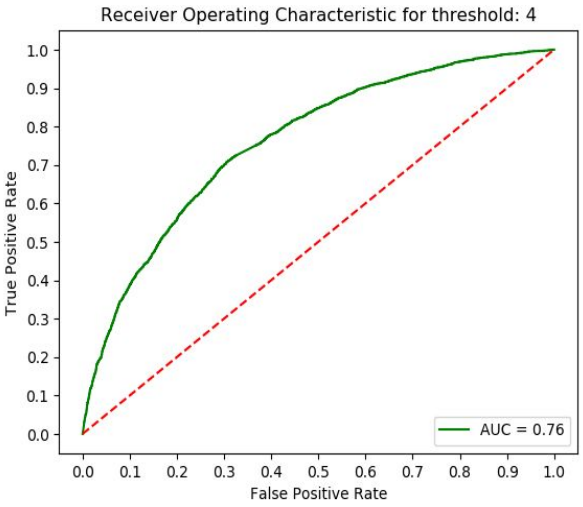
Minimum Average RMSE	1.5832288392820195

Q15) Plot the ROC curves for the k-NN collaborative filter designed in question 10 for threshold values [2.5, 3, 3.5, 4]. For the ROC plotting use the k found in question 11. For each of the plots, also report the area under the curve (AUC) value.









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.

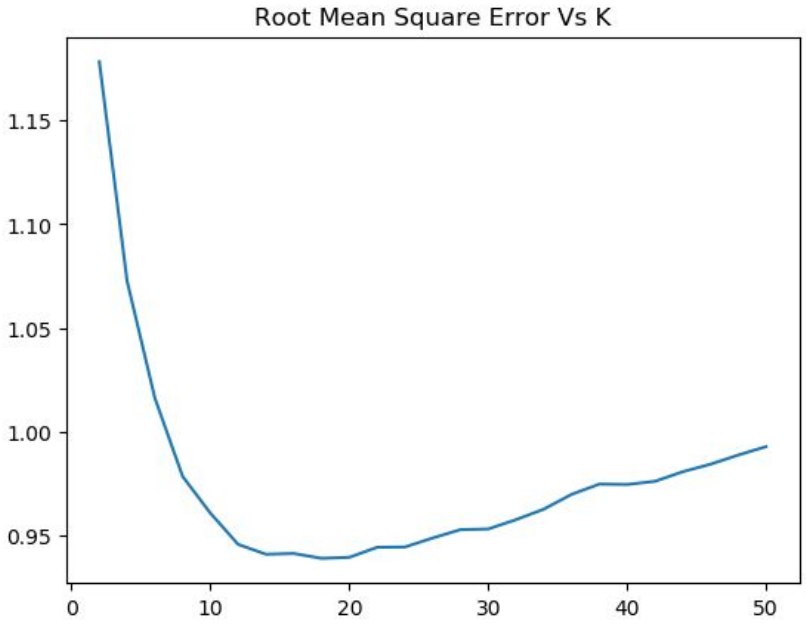
Since both U and V are unknowns, Equation 5 is not convex. However, if we fix one of the unknowns, the optimization problem becomes quadratic and can be solved optimally. We can either fix U or V.

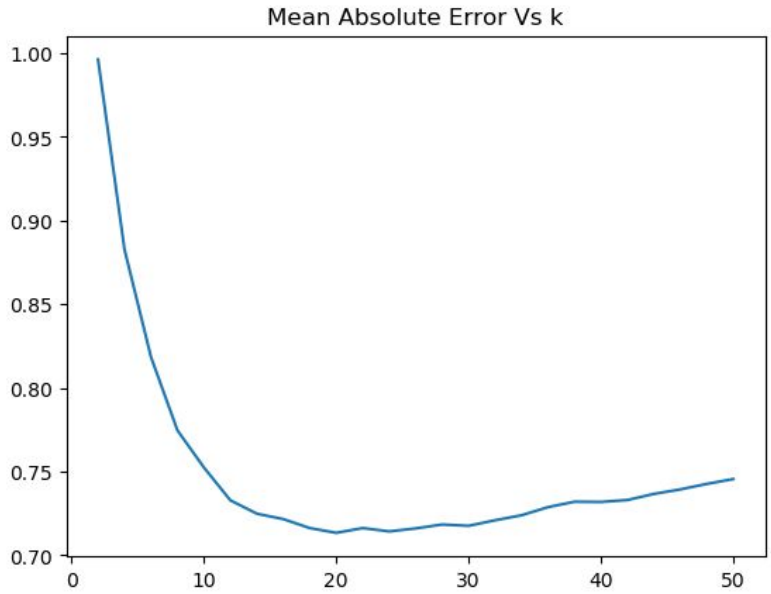
Keeping 'U' fixed, we solve for each of the n rows of 'V' by treating the problem as a least-squares model in each case. Let $\overline{v_j}$ be the jth row of V. In order to determine the optimal vector $\overline{v_j}$, we wish to minimize the following equation.

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

It's a least-squares regression problem in $v_{j1},\ v_{j2}, \ldots.$ The terms $u_{i1},\ u_{i2}, \ldots$ are treated as constants and $v_{j1},\ v_{j2}, \ldots.$ are treated as optimization variables.

Q17) Design a NNMF-based collaborative filter to predict the ratings of the movies in the MovieLens dataset and evaluate it's performance using 10-fold cross-validation. Plot the average RMSE (Y-axis) against k (X-axis) and the average MAE (Y-axis) against k (X-axis). For solving this question, use the default value for the regularization parameter.



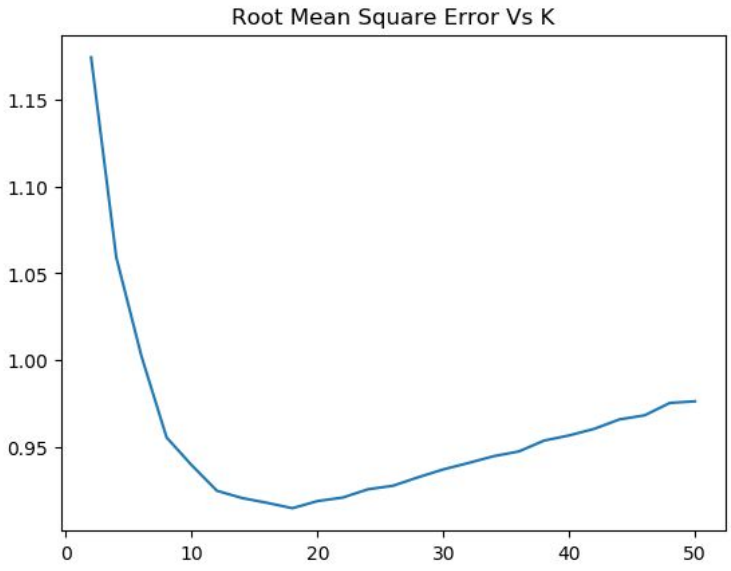


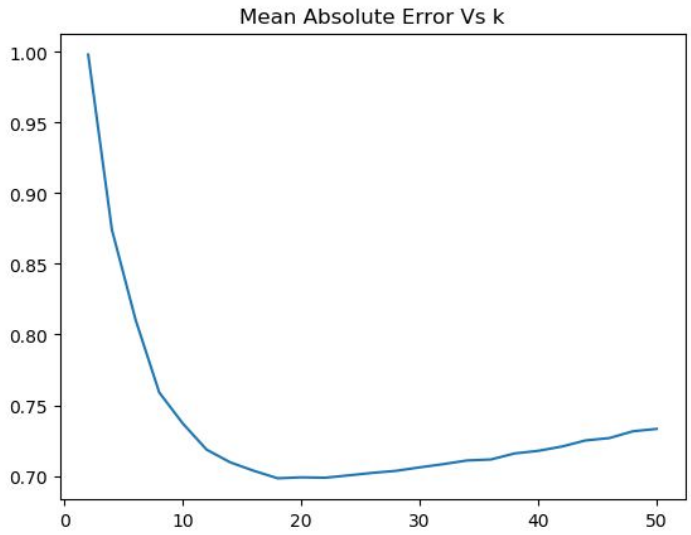
Q18) Use the plot from question 17, to find the optimal number of latent factors. Optimal number of latent factors is the value of k that gives the minimum average RMSE or the minimum average MAE. Please report the minimum average RMSE and MAE. Is the optimal number of latent factors same as the number of movie genres?

Optimal number of latent factors	20
Minimum Average RMSE	0.938948991763
Minimum Average MAE	0.713393470868

Yes, the number of genres (19) is almost equal to the number of latent factors.

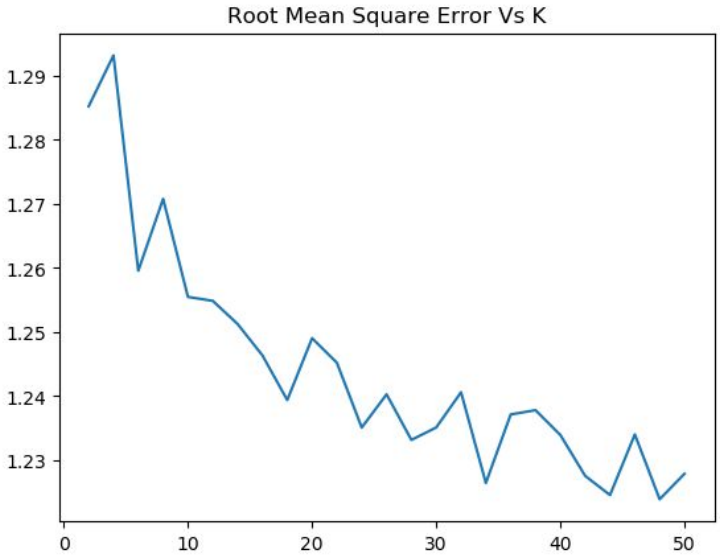
Q19) Design a NNMF collaborative filter to predict the ratings of the movies in the popular movie trimmed test set and evaluate it's performance using 10-fold cross validation. Plot average RMSE (Y-axis) against k (X-axis). Also, report the minimum average RMSE

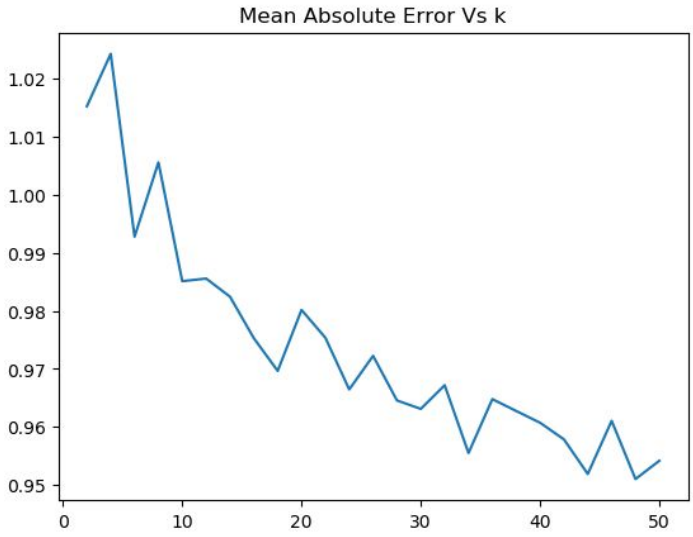




Minimum Average RMSE	0.9148999216656902

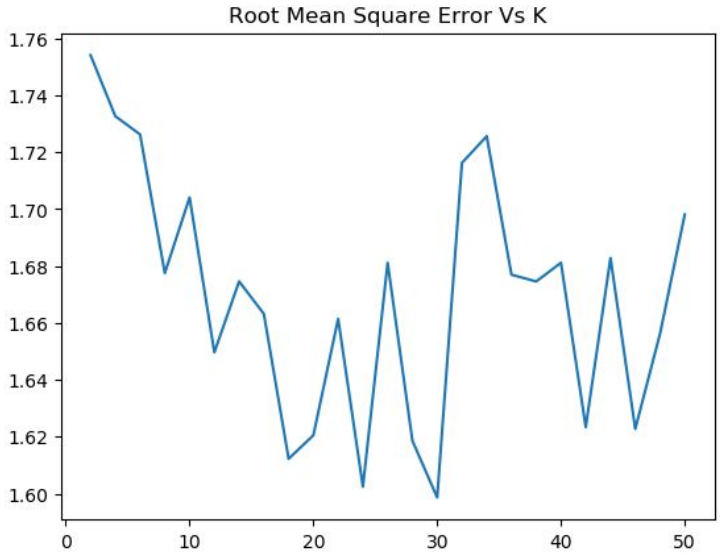
Q20) Design a NNMF collaborative filter to predict the ratings of the movies in the unpopular movie trimmed test set and evaluate it's performance using 10-fold cross validation. Plot average RMSE (Y-axis) against k (X-axis). Also, report the minimum average RMSE

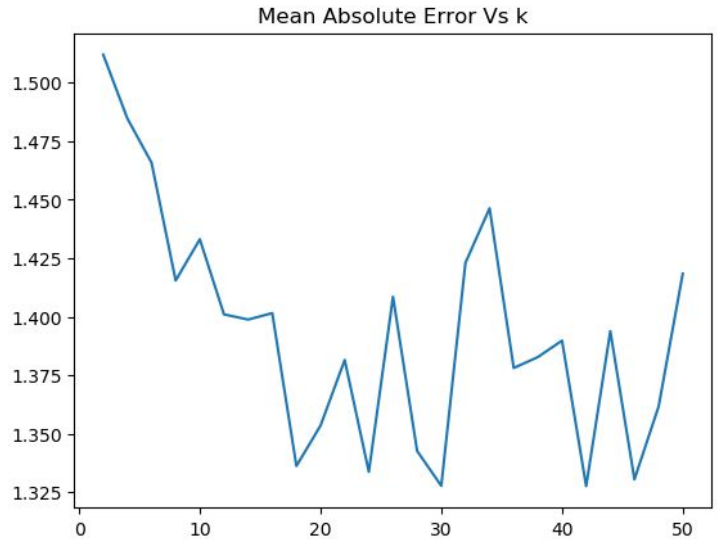




Minimum Average RMSE	1.2239248216322236

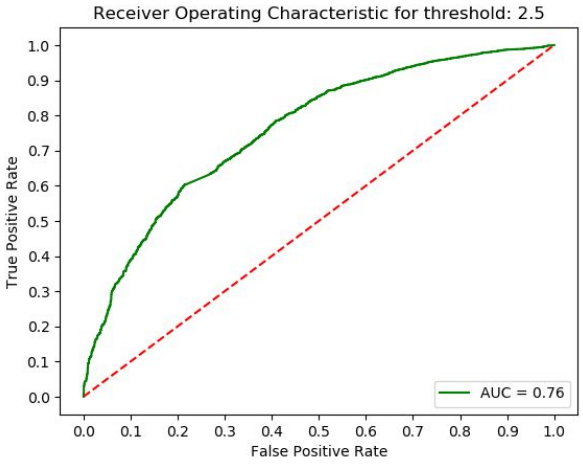
Q21) Design a NNMF collaborative filter to predict the ratings of the movies in the high variance movie trimmed test set and evaluate it's performance using 10-fold cross validation. Plot average RMSE (Y-axis) against k (X-axis). Also, report the minimum average RMSE

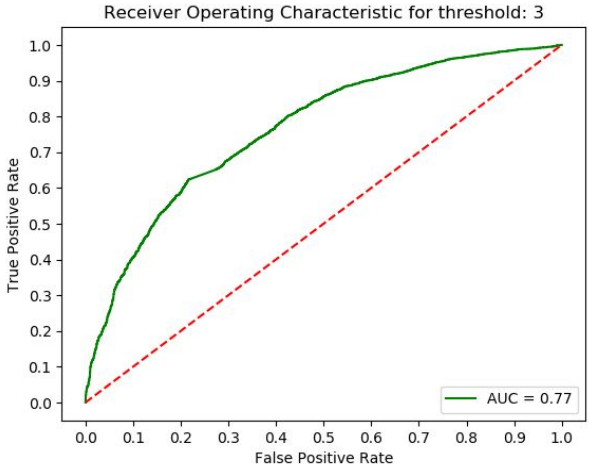


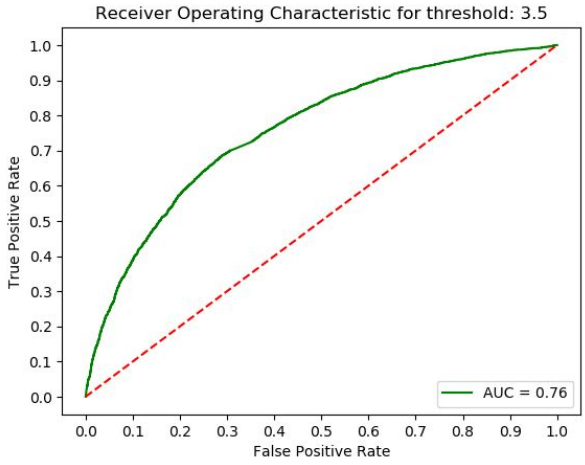


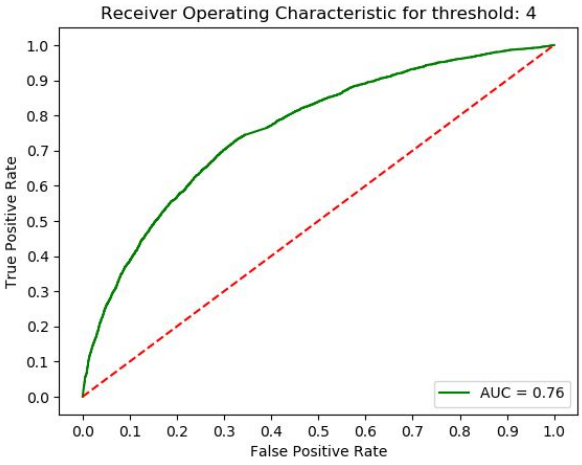
Minimum Average RMSE	1.5986459466801832

Q22) Plot the ROC curves for the NNMF-based collaborative filter designed in question 17 for threshold values [2.5,3,3.5,4]. For the ROC plotting use the optimal number of latent factors found in question 18. For each of the plots, also report the area under the curve (AUC) value.









Q23) Perform Non-negative matrix factorization on the ratings matrix R to obtain the factor matrices U and V , where U represents the user-latent factors interaction and V represents the movie-latent factors interaction (use k = 20). For each column of V , sort the movies in descending order and report the genres of the top 10 movies. Do the top 10 movies belong to a particular or a small collection of genre? Is there a connection between the latent factors and the movie genres?

In the following table, the columns are the latent factors and the values in it represent the counts.

	0	1	2	3	4	5	7
Comedy	5	4	6	4	7	4	3
Drama	4	6	5	7	3	5	3
Action	4		4	2	3	6	2
War	3		2				
Adventure	3	3	3	2	2	4	
Sci-Fi	3	3			2	4	5
IMAX	3					4	
Mystery	2		2		2	2	
Romance	2	3	3	2	4	2	
Crime	2		2	2	2	3	
Fantasy	2	2	2			2	
Documentary	2	2		2	3		
Children	2			2	2		
Thriller	2		4	2	3	4	2
Western	2						
Animation		2					2
Musical		2			2		2

Horror		2				3	4

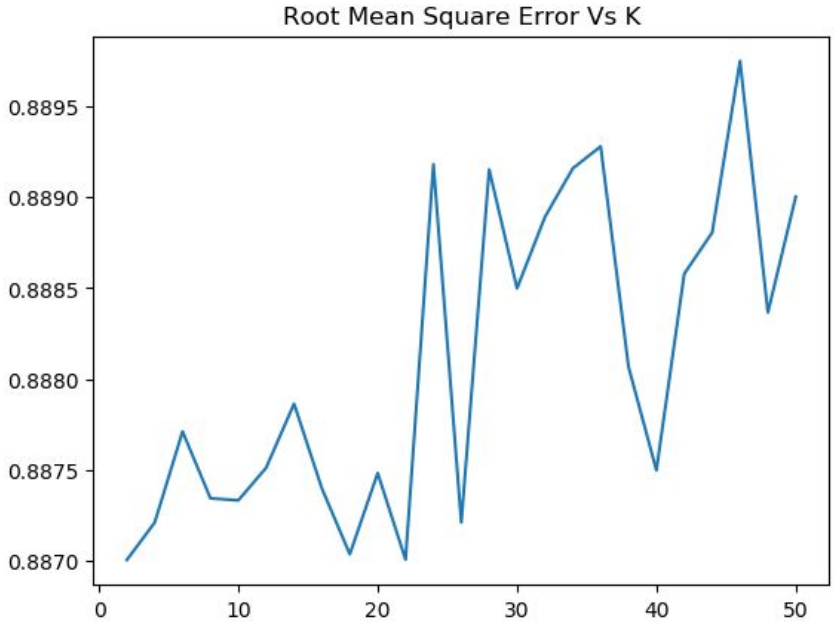
The below table shows the genres for the top 10 movies from which the above counts have been calculated.

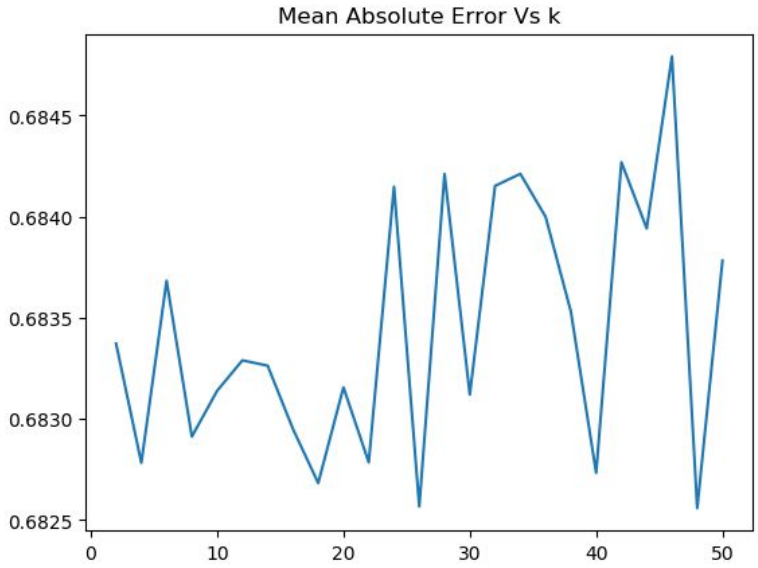
	Genres
0	['Drama', 'Mystery', 'Romance'], ['Action', 'Comedy', 'Crime', 'Fantasy'], ['Comedy', 'Documentary'], ['Drama', 'War'], ['Children', 'Comedy'], ['Action', 'Adventure', 'Sci-Fi', 'War', 'IMAX'], ['Thriller'] ['Comedy', 'Western'], ['Action', 'Adventure', 'Sci-Fi', 'IMAX'], ['Drama']
1	['Comedy', 'Documentary'], ['Adventure', 'Animation'], ['Musical'] ['Comedy', 'Romance'], ['Drama', 'Fantasy', 'Horror'], ['Comedy'], ['Drama', 'Sci-Fi'], ['Drama'], ['Adventure', 'Drama', 'Sci-Fi'], ['Drama', 'Romance']
2	['Adventure', 'Drama', 'Fantasy', 'Romance'], ['Action', 'Adventure', 'Drama', 'Thriller'], ['Action', 'War'], ['Action', 'Crime', 'Thriller'], ['Comedy'] ['Comedy', 'Drama', 'Romance'], ['Comedy'], ['Comedy', 'Mystery', 'Thriller'] ['Drama'], ['Comedy']
3	['Drama'], ['Comedy', 'Crime'], ['Comedy', 'Drama'], ['Documentary', 'Drama'], ['Action'], ['Drama', 'Thriller'], ['Adventure', 'Children'], ['Drama', 'Romance'] ['Comedy'], ['Drama']
4	['Comedy', 'Crime', 'Mystery', 'Thriller'], ['Action', 'Comedy'], ['Documentary'] ['Action', 'Adventure', 'Sci-Fi', 'Thriller'], ['Comedy', 'Drama'], ['Drama', 'Romance'], ['Children', 'Comedy', 'Musical', 'Romance'], ['Comedy'] ['Documentary'], ['Comedy', 'Romance']
7	['Action'], ['Drama'], ['Animation'], ['Thriller'], ['Drama', 'Sci-Fi'], ['Horror', 'Sci-Fi'], ['Comedy', 'Horror', 'Sci-Fi'], ['Horror'], ['Comedy', 'Musical'], ['Sci-Fi']

Observing the count of each genre across all the 20 latent factors, we see that under a majority of latent factors, Comedy and Drama have a higher count. So we can conclude that the top 10 movies indeed belong to small collection of genre.

Observing the genre with the highest count in each column corresponding to a latent factor, eg. Comedy has high count of 7 in column 4, Drama has count of 7 in column 3 while Sci-Fi has count of 5 in column 7, we can see that there is a connection between the latent factors and the movie genre in a sense that a latent factor is mapping to the aspects of a particular genre like how dramatic a movie is, how comedy a movie is.

Q24) Design a MF with bias collaborative filter to predict the ratings of the movies in the MovieLens dataset and evaluate it's performance using 10-fold cross-validation. Plot the average RMSE (Y-axis) against k (X-axis) and the average MAE (Y-axis) against k (X-axis). For solving this question, use the default value for the regularization parameter.

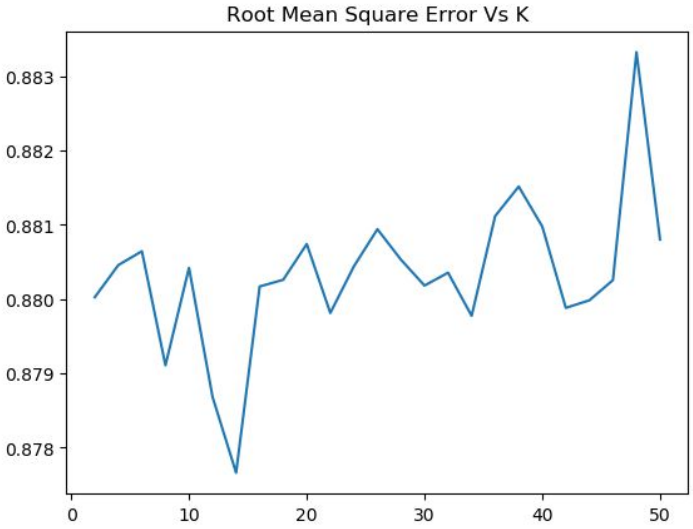


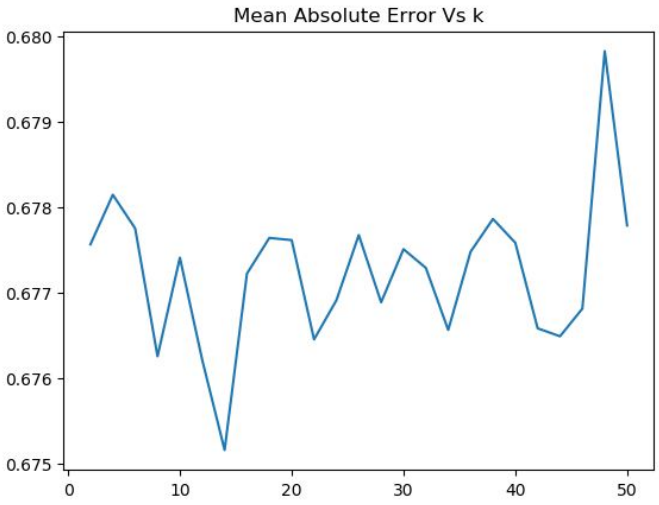


Q25) Use the plot from question 24, to find the optimal number of latent factors. Optimal number of latent factors is the value of k that gives the minimum average RMSE or the minimum average MAE. Please report the minimum average RMSE and MAE.

Optimal number of latent factors	2
Minimum Average RMSE	0.887005061726
Minimum Average MAE	0.68255951152

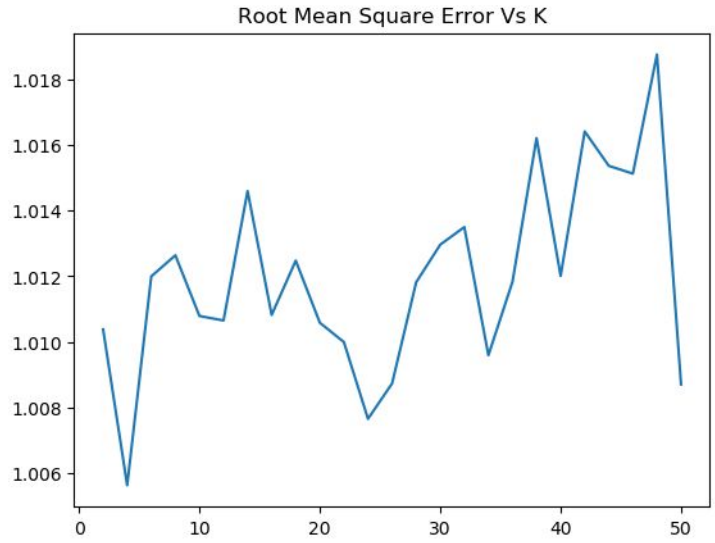
Q26) Design a MF with bias collaborative filter to predict the ratings of the movies in the popular movie trimmed test set and evaluate it's performance using 10-fold cross validation. Plot average RMSE (Y-axis) against k (X-axis). Also, report the minimum average RMSE

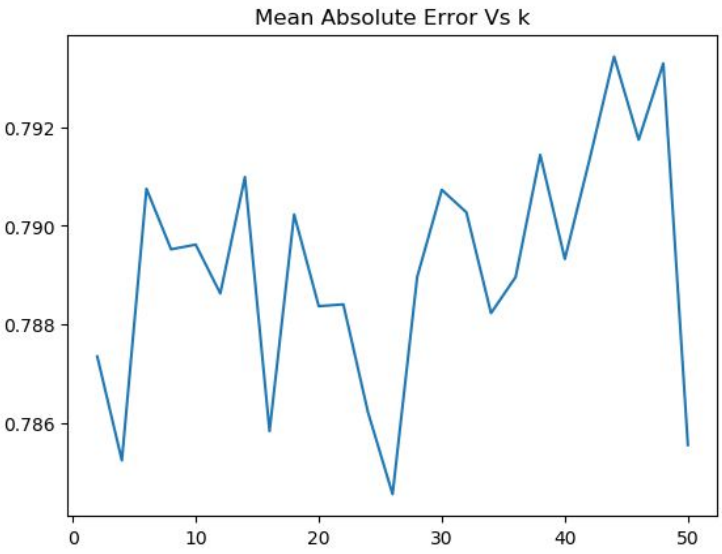




Minimum Average RMSE	0.8776621218344965

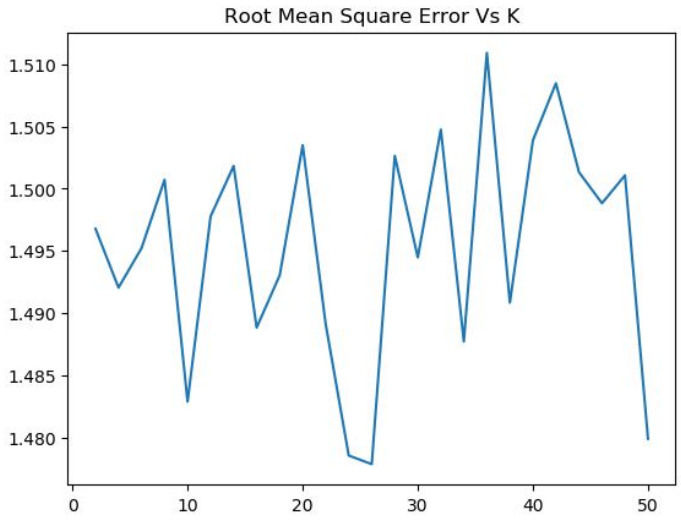
Q27) Design a MF with bias collaborative filter to predict the ratings of the movies in the unpopular movie trimmed test set and evaluate it's performance using 10-fold cross validation. Plot average RMSE (Y-axis) against k (X-axis). Also, report the minimum average RMSE

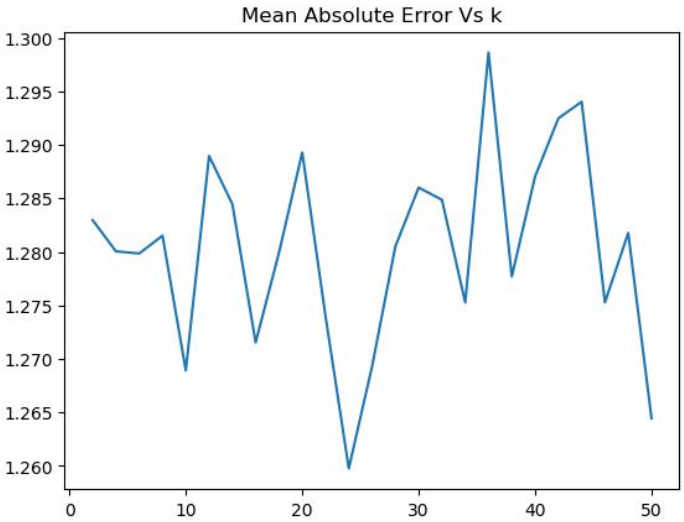




Minimum Average RMSE	1.005634065205908

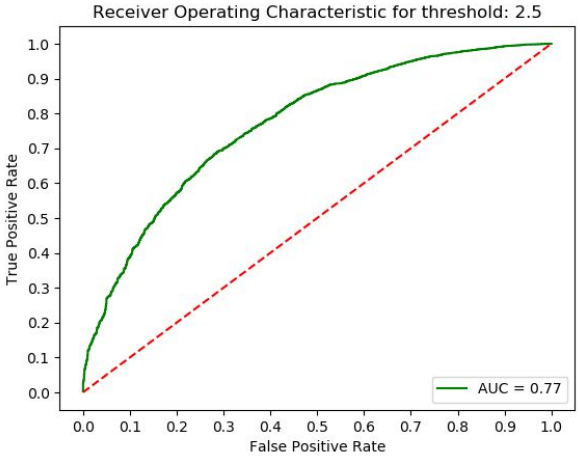
Q28) Design a MF with bias collaborative filter to predict the ratings of the movies in the high variance movie trimmed test set and evaluate it's performance using 10-fold cross validation. Plot average RMSE (Y-axis) against k (X-axis). Also, report the minimum average RMSE

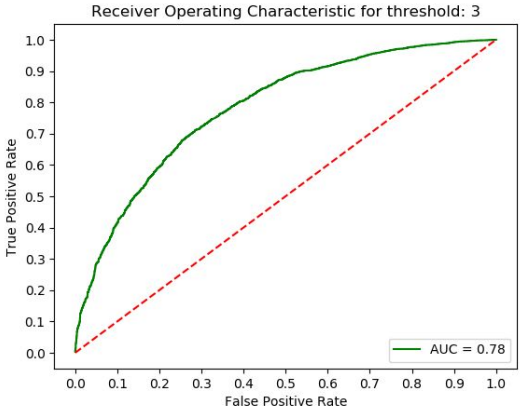


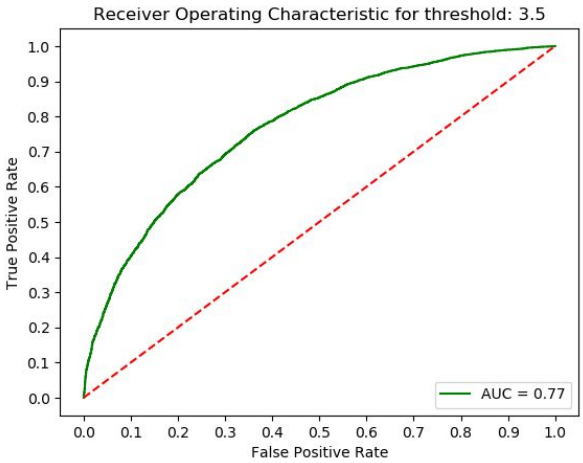


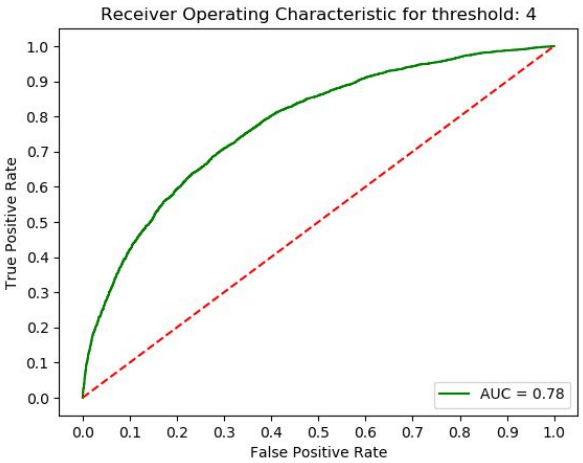
Minimum Average RMSE	1.4778784359706303

Q29) Plot the ROC curves for the MF with bias collaborative filter designed in question 24 for threshold values [2.5,3,3.5,4]. For the ROC plotting use the optimal number of latent factors found in question 25. For each of the plots, also report the area under the curve (AUC) value.









Q30) Design a naive collaborative filter to predict the ratings of the movies in the MovieLens dataset and evaluate its performance using 10-fold cross validation. Compute the average RMSE by averaging the RMSE across all 10 folds. Report the average RMSE.

Average RMSE	0.95686922279104059
Average MAE	0.74578669241395756

Q31) Design a naive collaborative filter to predict the ratings of the movies in the popular movie trimmed test set and evaluate it's performance using 10-fold cross validation. Compute the average RMSE by averaging the RMSE across all 10 folds. Report the average RMSE.

Average RMSE	0.95367084033646576
Average MAE	0.74416909615146321

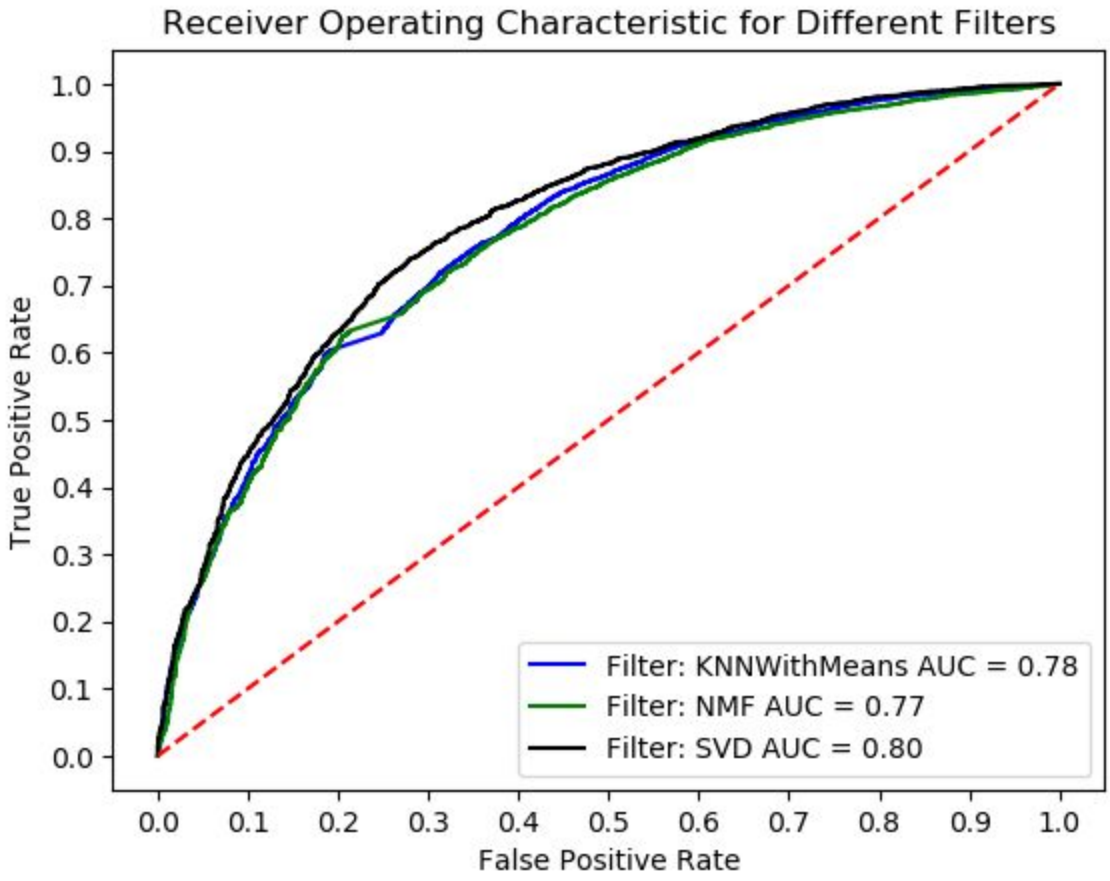
Q32) Design a naive collaborative filter to predict the ratings of the movies in the unpopular movie trimmed test set and evaluate it's performance using 10-fold cross validation. Compute the average RMSE by averaging the RMSE across all 10 folds. Report the average RMSE.

Average RMSE	1.0108116609797442
Average MAE	0.77322056806945016

Q33) Design a naive collaborative filter to predict the ratings of the movies in the high variance movie trimmed test set and evaluate it's performance using 10-fold cross validation. Compute the average RMSE by averaging the RMSE across all 10 folds. Report the average RMSE.

Average RMSE	1.5230540579919989
Average MAE	1.2633844105845642

Q34) Plot the ROC curves (threshold = 3) for the k-NN, NNMF, and MF with bias based collaborative filters in the same figure. Use the figure to compare the performance of the filters in predicting the ratings of the movies.

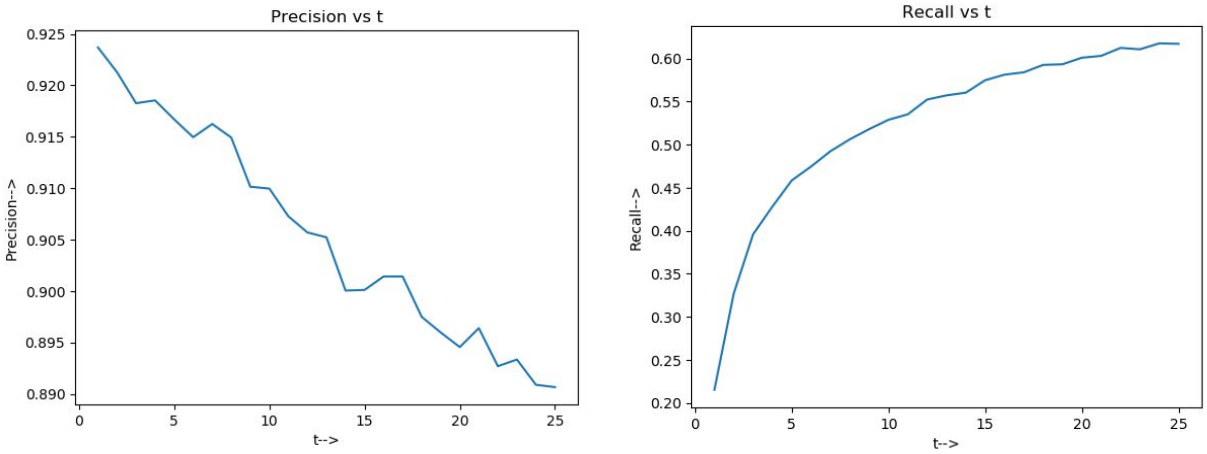


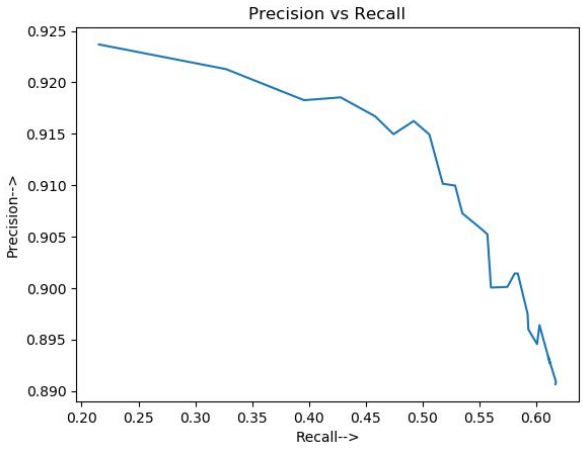
Q35) Precision and Recall are defined by the mathematical expressions given by equations 12 and 13 respectively. Please explain the meaning of precision and recall in your own words.

Precision: Out of all the items recommended to the user how many are actually liked by that user.

Recall: Out of all the items liked by the user, how many were recommended to that user.

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). Use the k found in question 11 and sweep t from 1 to 25 in step sizes of 1. For each plot, briefly comment on the shape of the plot.

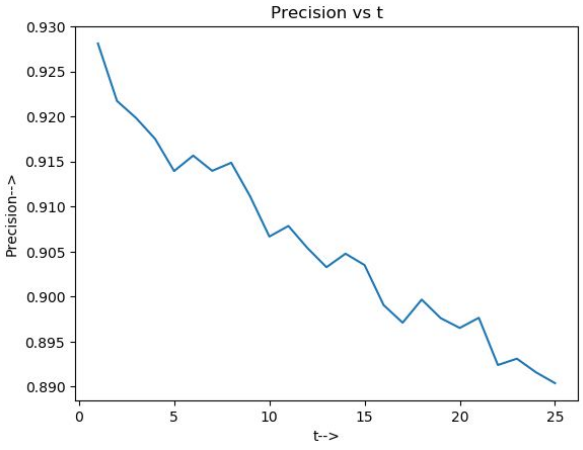


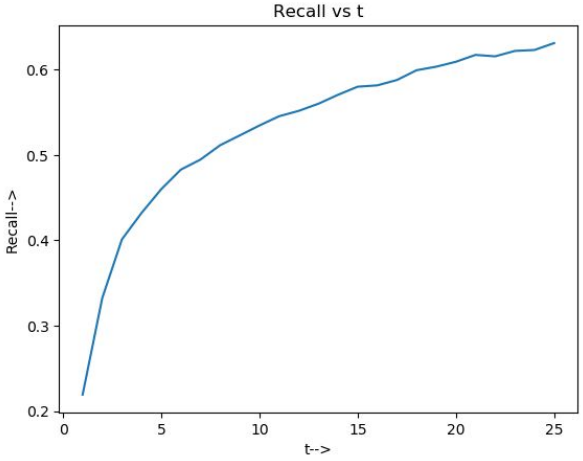


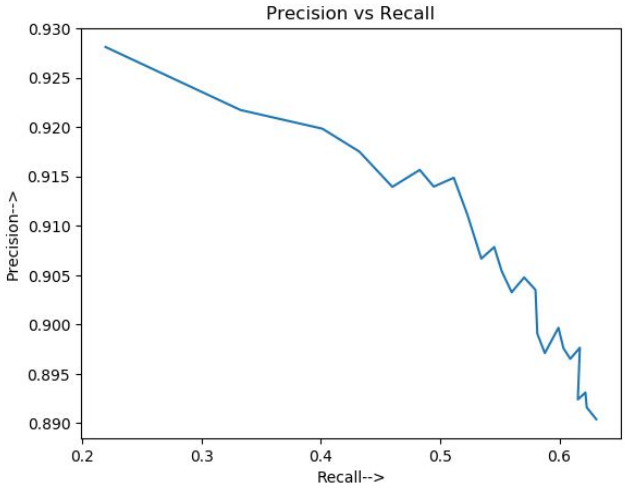
Precision vs t	As ‘t’ increases, the denominator for precision, S(t) increases and hence the precision decreases
Recall vs t	The denominator |G| remains constant and as ‘t’ increases the numerator i.e. intersection between items recommended and ground truth items may increase and hence the recall increases
Precision vs Recall	Based on the above 2 reasonings, we can observe that as recall

	increases, precision decreases.

Q37) Plot average precision (Y-axis) against t (X-axis) for the ranking obtained using NNMF-based collaborative filter predictions. Also, plot the average recall (Y-axis) against t (X-axis) and average precision (Y-axis) against average recall (X-axis). Use optimal number of latent factors found in question 18 and sweep t from 1 to 25 in step sizes of 1. For each plot, briefly comment on the shape of the plot.



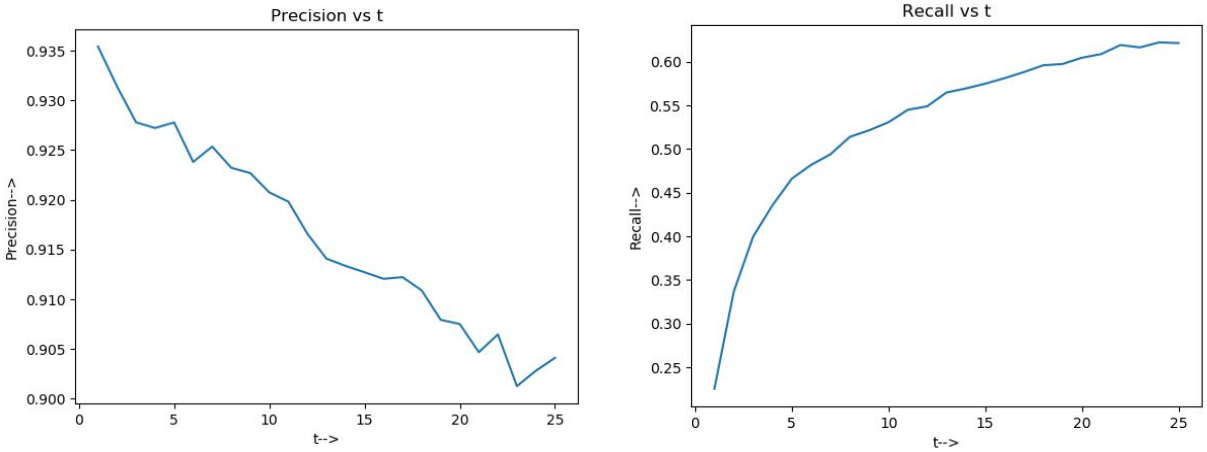


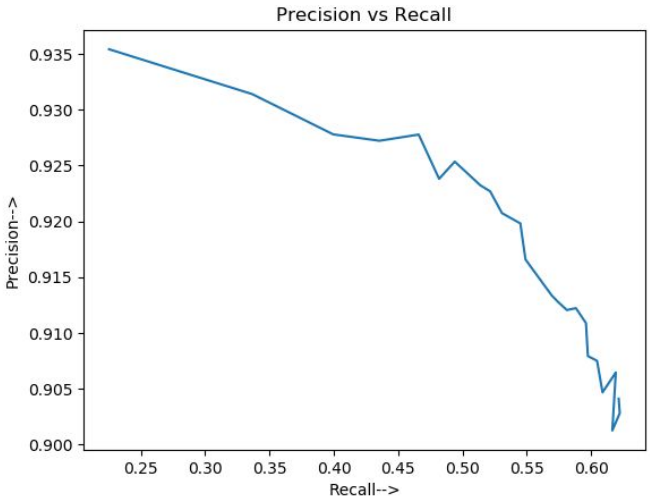


Precision vs t	As ‘t’ increases, the denominator for precision, S(t) increases and hence the precision decreases

Recall vs t	The denominator |G| remains constant and as 't' increases the numerator i.e. intersection between items recommended and ground truth items may increase and hence the recall increases
Precision vs Recall	Based on the above 2 reasonings, we can observe that as recall increases, precision decreases.

Q38) Plot average precision (Y-axis) against t (X-axis) for the ranking obtained using MF with bias-based collaborative filter predictions. Also, plot the average recall (Y-axis) against t (X-axis) and average precision (Y-axis) against average recall (X-axis). Use optimal number of latent factors found in question 25 and sweep t from 1 to 25 in step sizes of 1. For each plot, briefly comment on the shape of the plot.





Precision vs t	As ‘t’ increases, the denominator for precision, S(t) increases and hence the precision decreases
Recall vs t	The denominator |G| remains constant and as ‘t’ increases the numerator i.e. intersection between items recommended and ground truth items may increase and hence the recall increases
Precision vs Recall	Based on the above 2 reasonings, we can observe that as recall increases, precision decreases.

Q39) Plot the precision-recall curve obtained in questions 36,37, and 38 in the same figure. Use this figure to compare the relevance of the recommendation list generated using k-NN, NNMF, and MF with bias predictions.