EE219 Project3
Collaborative Filtering

Qidi Sang 705028670
Hui Wang 205036597
Zhonglin Zhang 005030520

1. Introduction

The basic models for recommender systems works with two kinds of data:
1. User-Item interactions such as ratings
2. Attribute information about the users and items such as textual profiles or relevant keywords

Models that use type 1 data are referred to as collaborative filtering methods, whereas models that use type 2 data are referred to as content based methods. In this project, we built recommendation system using collaborative filtering methods.

2. Collaborative Filtering Models

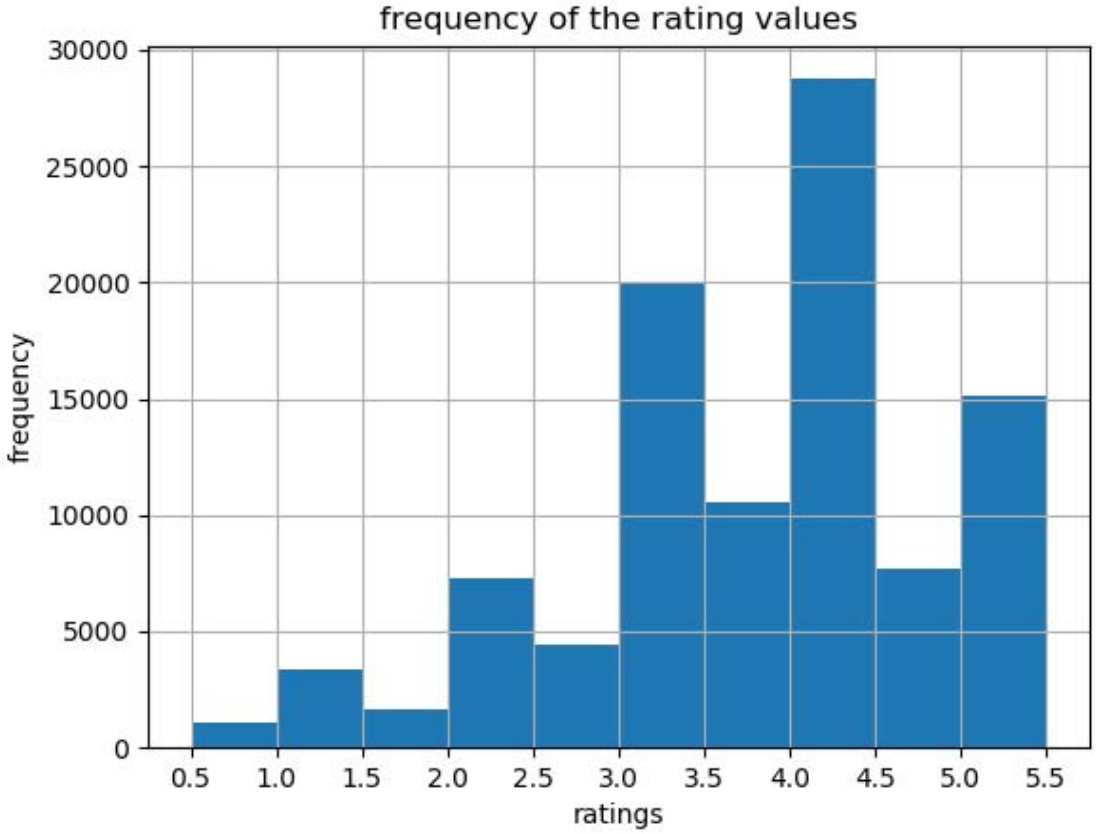
The basic idea of collaborative filtering methods is that these unspecified ratings can be imputed because the observed ratings are often highly correlated across various users and items. For example, consider two users named John and Molly, who have very similar tastes. If the ratings, which both have specified, are very similar, then their similarity can be identified by the filtering algorithm. In such cases, it is very likely that the ratings in which only one of them has specified a value, are also likely to be similar. This similarity can be used to make inferences about incompletely specified values. Most of the collaborative filtering methods focuses on leveraging either inter-item correlations or inter-user correlations for the prediction process.

In this project, we will implement and analyze the performance of two types of collaborative filtering methods:
1. Neighborhood-based collaborative filtering
2. Model-based collaborative filtering

3. MovieLens Dataset

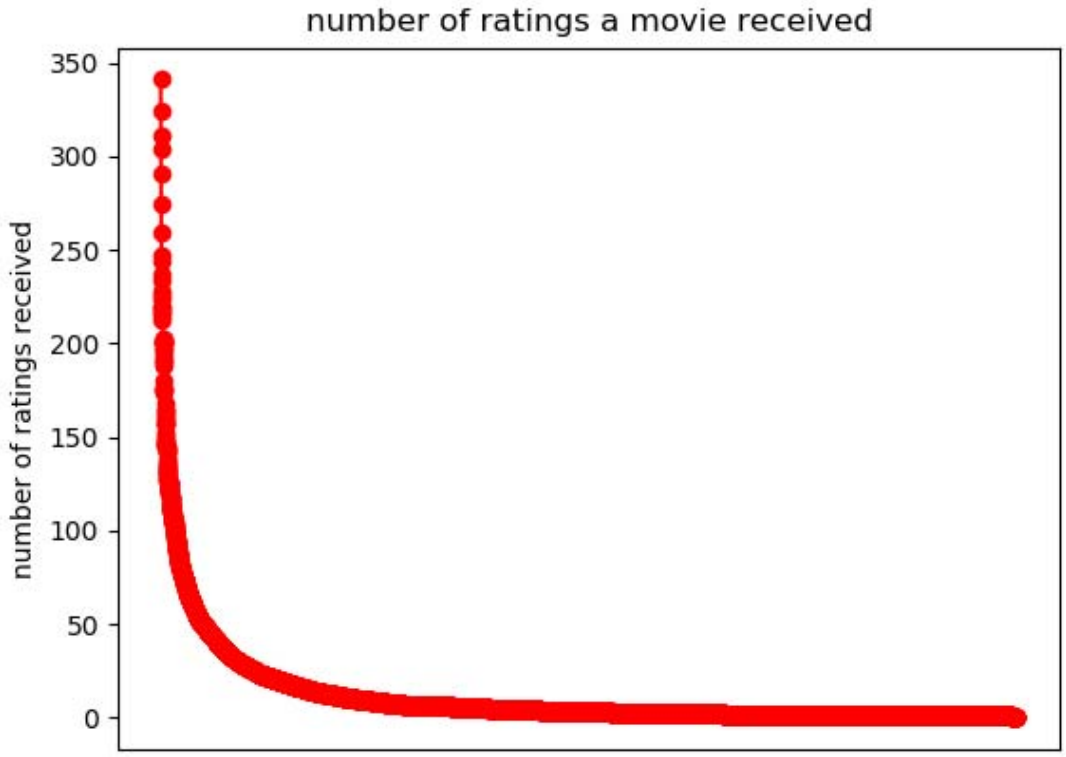
Question 1
Sparsity is 0.01633285017250883

Question 2

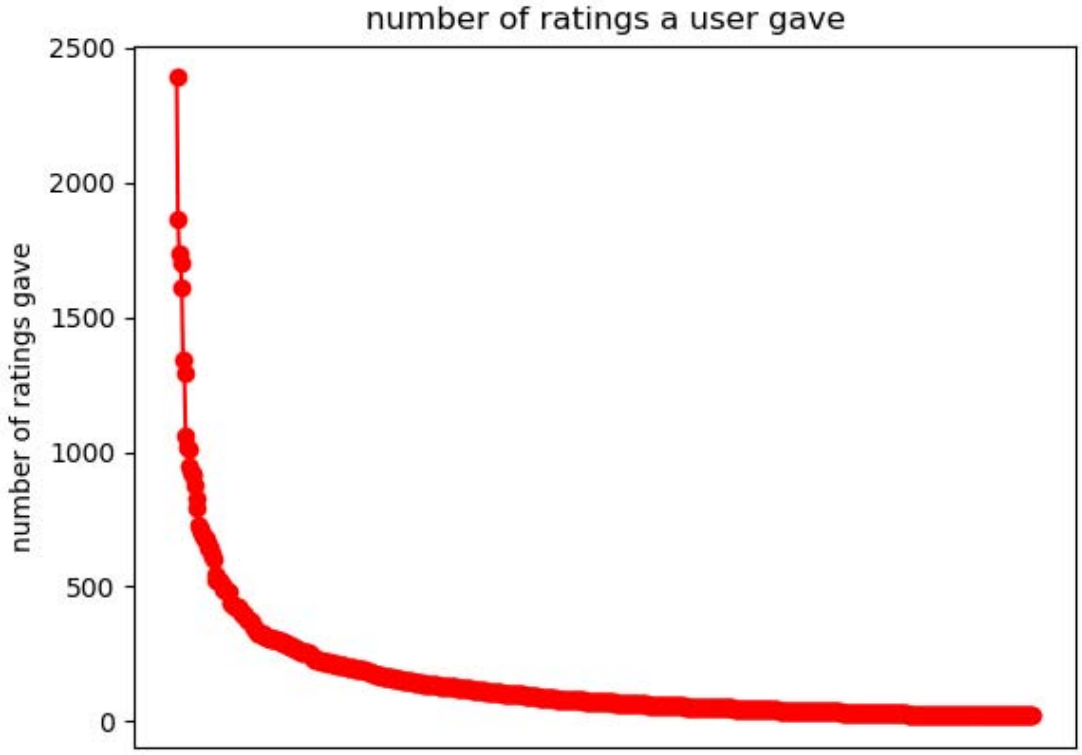


We can see the integer ratings (1,2,3,4,5) seem to appear more than adjacent decimal ratings; and most ratings appear in range of 3 to 4, they make up of roughly 60% of whole ratings.

Question 3


number of ratings a movie received
number of ratings received
350
300
250
200
150
100
50
0

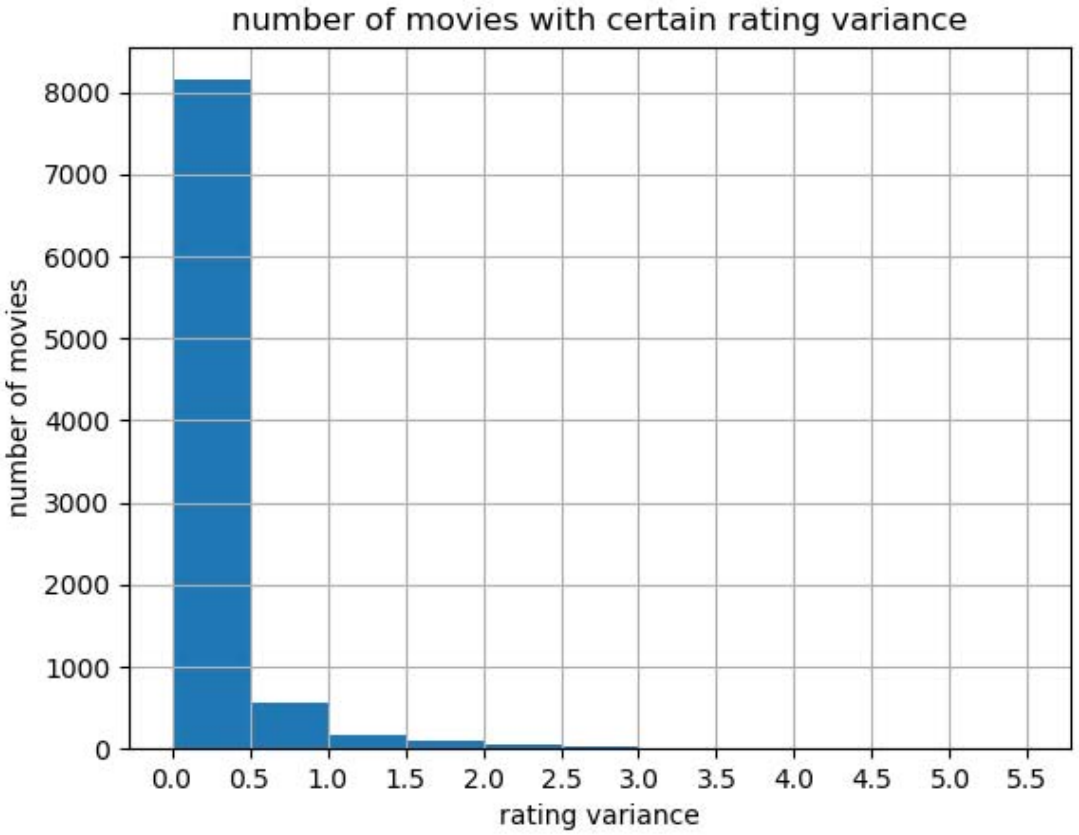
Question 4



Question 5

The number of ratings a movie received falls drastically and it has a very long tail, which means that most of the movies received less than 20 ratings. Then in the recommendation process, we are going to deal with a very sparse matrix. Therefore, maybe we should focus on the data points that are given to us rather than operate on this sparse matrix.

Question 6



The movies that has variance less than 0.5 is the absolute majority (about 90%)of all movies.

4. Neighborhood-based Collaborative Filtering

Question 7

$$\mu_u = \frac{\sum_{k \in I_u} r_{uk}}{card I_u}$$

p.s. card I_u means the number of elements in set I_u

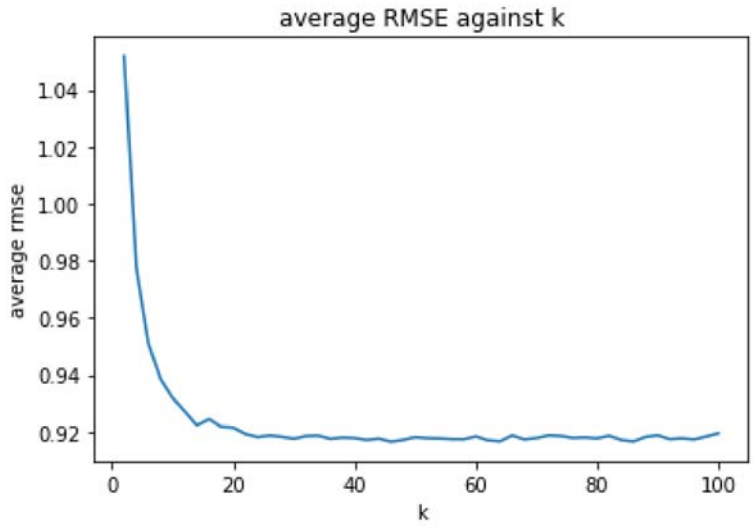
Question 8

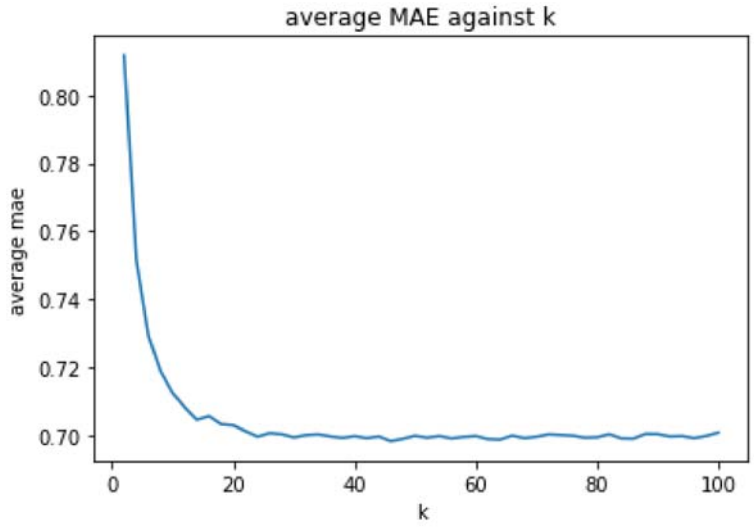
$I_u \cap I_v$ means the set of item indices for which ratings have been specified by both user u and user v. In other words, it's the movies' indices for which both user u and user v have given their ratings. It's possible that $I_u \cap I_v = \emptyset$, since the rating matrix R is sparse and contains many zeros. And if r_{uk} is zero, it means that movie k has not been rated by user u. Or in other words, there exists the possibility that user u and user v rate two sets of completely different movies.

Question 9

Different users have their own degrees of rating, which are different standards according to each user. Tough users usually give low ratings and easy users always give high ratings. And this phenomenon will impact the accuracy of prediction. By subtracting the average ratings, ratings will be established on a uniform standard, so that the predictions will be less influenced and more objective.

Question 10



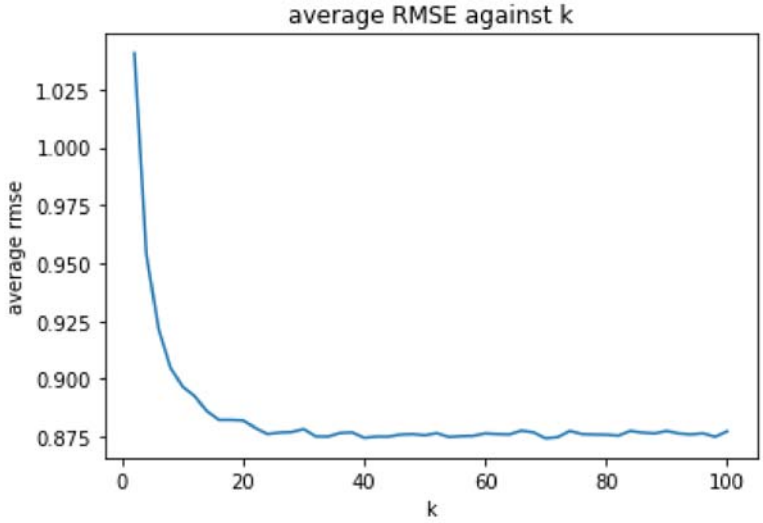


In this question we used KNNWithMeans from the surprise tool and swept k (number of neighbors) from 2 to 100 in step sizes of 2, and for each k we calculated the average RMSE and average MAE. Above two plots are average RMSE against k and average MAE against k. From the plots we can see that when k is small (less than 20), both average RMSE and average MAE are high, and the scores decrease when k increases. But as k keeps increasing, the scores do

not have significant decreases. The average RMSE and average MAE converge to a steady-state value.

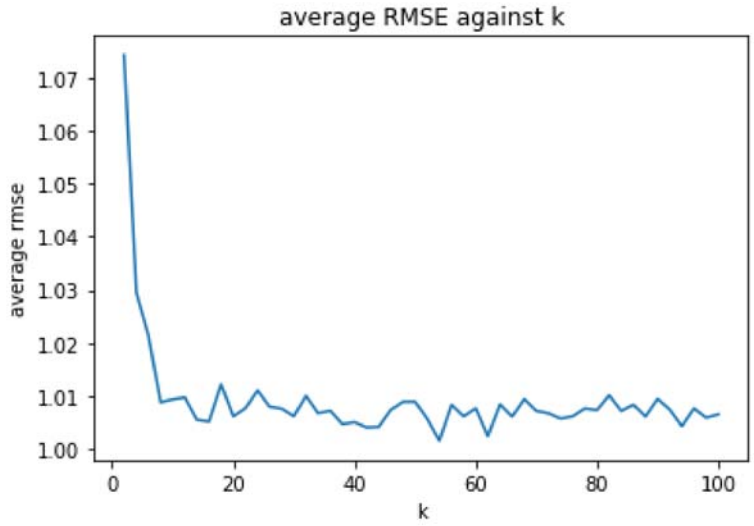
Question 11
From the plots in question 10, the minimum k values for both average RMSE and average RMSE are 30.

Question 12
The minimum average RMSE: 0.874351633166



In this question, we trimmed the test set to contain movies that has received more than 2 ratings. This means for each movie, there is sufficient information for the k-NN algorithm to predict and thus the prediction will be more accurate. From the plot above we can also see that the average RMSE is lower than that of the complete test set. The minimum average RMSE is 0.874351633166, which is the lowest among the three trimmed test set and means that this trimmed test set have the best prediction result.

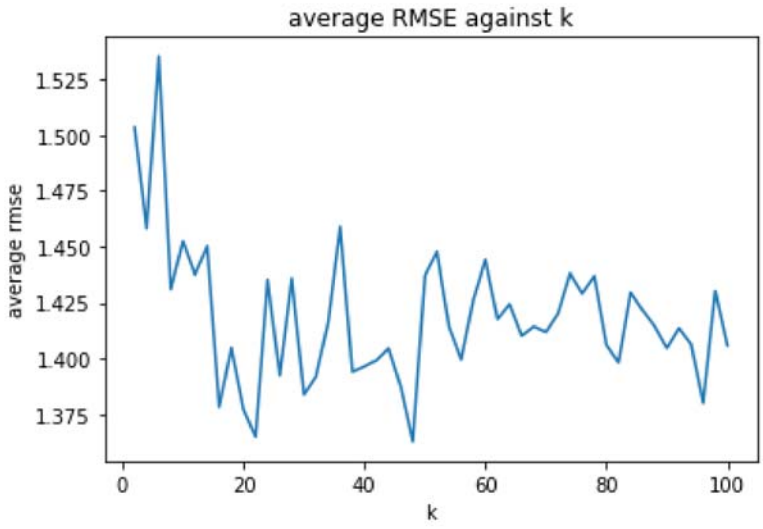
Question 13
The minimum average RMSE: 1.00157595559



In this question, we trimmed the test set to contain movies that has received less than or equal to 2 ratings. From the plot we can see that the average RMSE curve is not smooth. This is because these movies have less than 2 ratings so that there's no sufficient data for the k-NN algorithm to predict. Therefore,co the error rate is not stable and is relatively higher than the complete test set. The minimum average RMSE is 1.00157595559, which higher than popular-trimmed test set and means that this trimmed test set's prediction result is worse than popular-trimmed test set.

Question 14

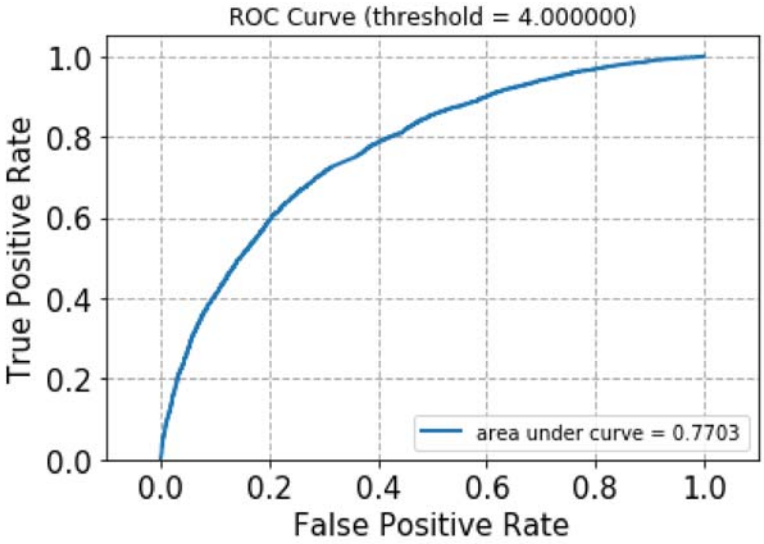
The minimum average RMSE: 1.36279344409

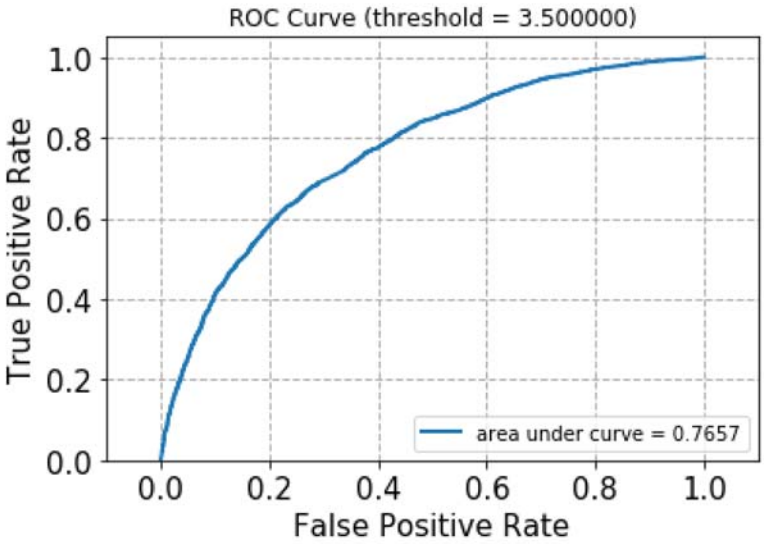


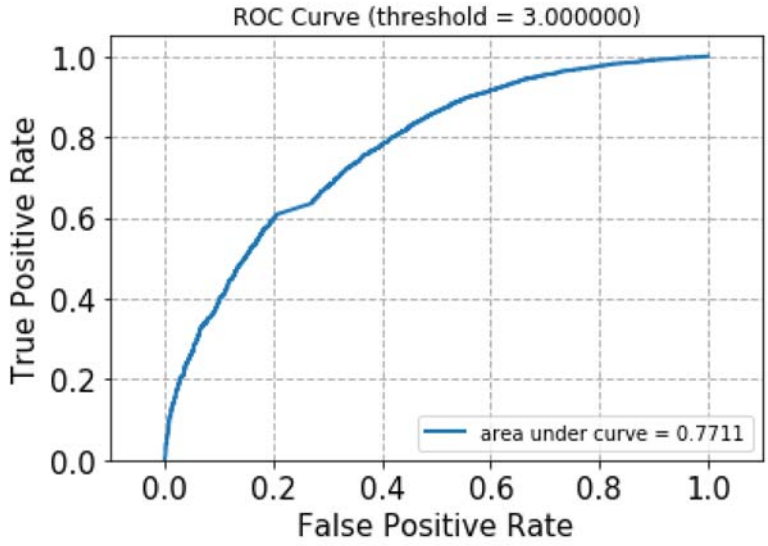
In this question, we trimmed the test set to contain movies that has variance (of the rating values received) of at least 2 and has received at least 5 ratings in the entire dataset. High variance, in other words, means that the evaluation of this movie is highly controversial, the ratings of different users on this movie vary widely. Thus based on such test set, the prediction

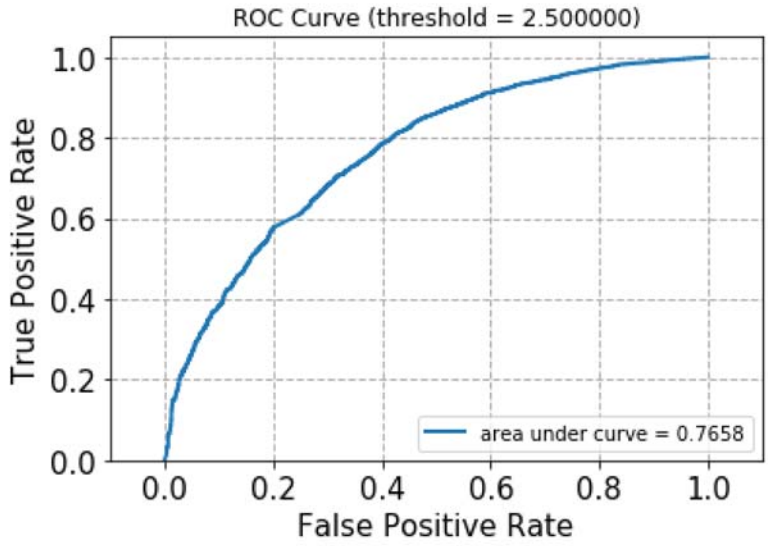
error rate will be relatively higher than un-trimmed test set. The minimum average RMSE is 1.36279344409, which is the highest among the three trimmed test set and this means that this trimmed test set have the worst prediction result.

Question 15









Above four plots are the ROC curve for the k-NN collaborative filter with the best k = 30 and using threshold values [2.5, 3, 3.5, 4]. From the four plots, we can see that there's only small difference between the AUC values. When threshold = 3, 4, the AUC values are relatively higher, which means better prediction result.

5. Model-based Collaborative Filtering

Question 16

The optimization problem in equation 5 is convex. We can define:

$$UV^T = M_{m \times n}$$

Then we can change the expression to:

$$G(M) = \sum_{i=1}^{m}\sum_{j=1}^{n} W_{ij}(r_{ij} - m_{ij})^2$$

$$\frac{\partial^2 G(M)}{\partial^2 m_{ij}} = 2W_{ij} \geq 0$$

So G(M) is a convex function.

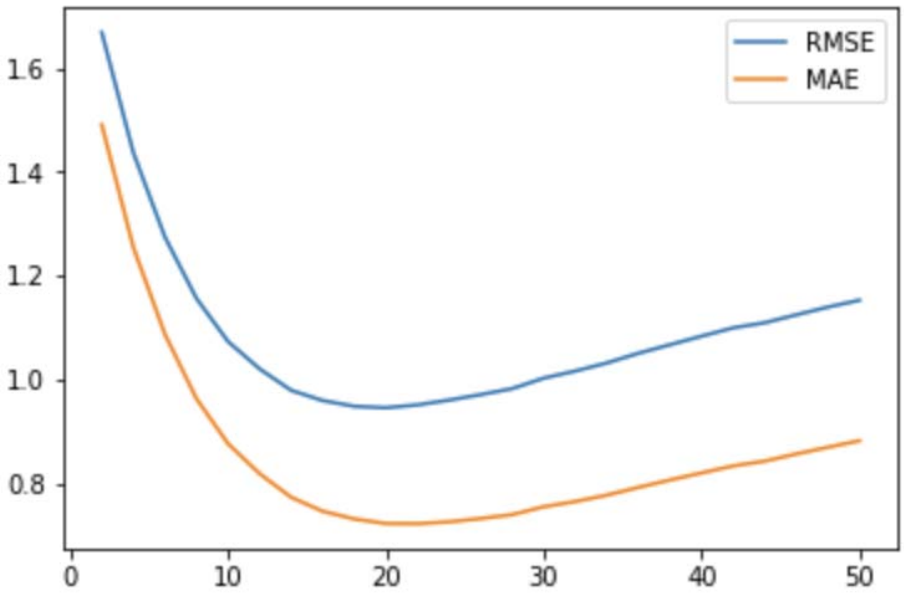
When U is fixed, we can consider it as a weight vector and the equation can be written as:

$$minimize \sum_{i,j}^{R} (r_{ij} - uV^T)$$

where R is the set of all the known ratings r_{ij}. In this way ,we can regard the optimization as a least-squares problem.

Question 17

In this question, we used NMF method in surprise package to build an NNMF-based collaborative filter. The number of latent factors k was swept from 2 to 50 and then we calculated the average RMSE and MAE using 10-fold cross-validation. The results were plotted in the following figure, in which Y-axis is RMSE/MAE measures and X-axis is k.



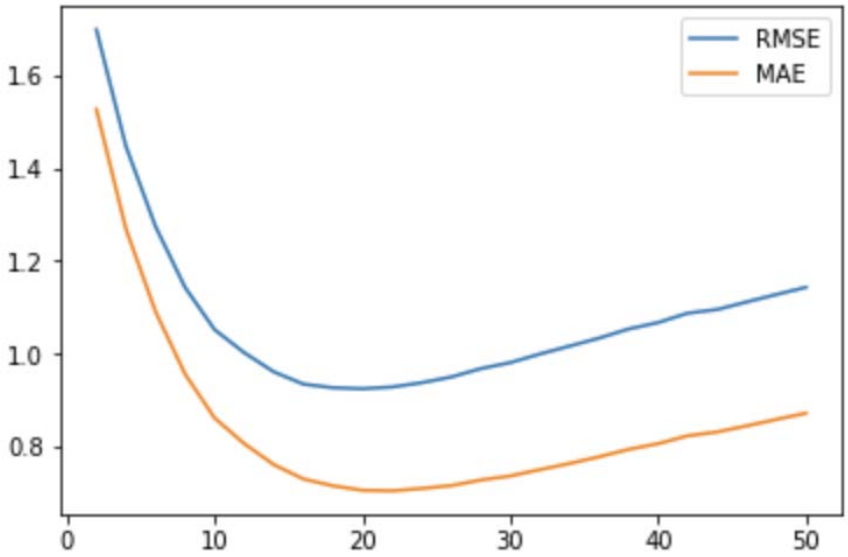
RMSE/MAE Plot

Question 18

According to the plot, the optimal number of latent factors for RMSE is 20 while the optimal number for MAE is 20. The average RMSE can reach its minimum value, 0.945, when k = 20, and the minimum of average MAE is 0.721 when k = 22. For the dataset we used in this project, there are 19 kinds of movies including 'no genres listed'. Hence, we can note that the optimal number of latent factors is quite close to the number of movie genres.

Question 19

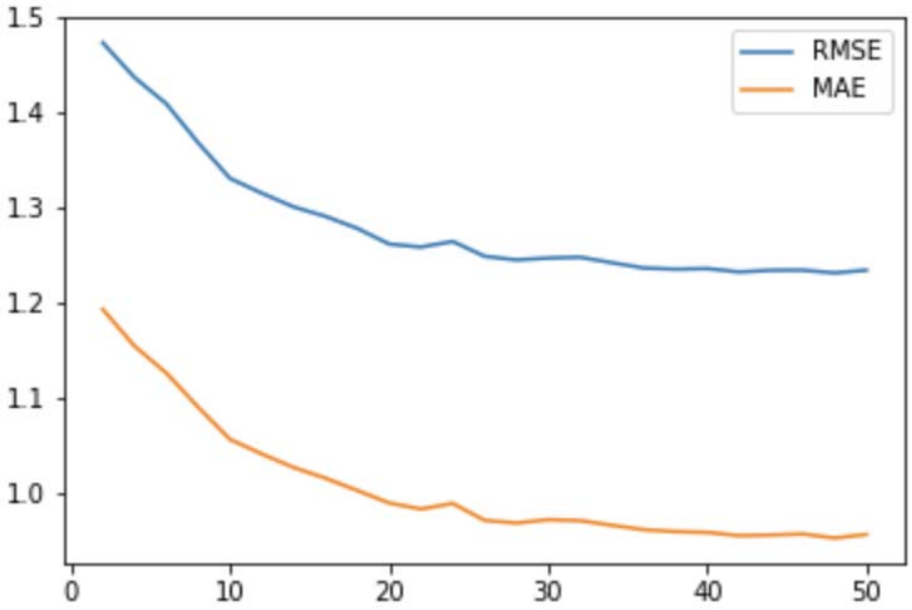
For this part, we need to trim the testset data before prediction and repeat the above steps. In the trimming function, we counted the number of ratings for each movie and picked the movies with more than 2 ratings as testset. The values of average RMSE/MAE against k are plotted in the following figure. The minimum average RMSE is 0.925 when k is 20.



RMSE/MAE Plot

Question 20

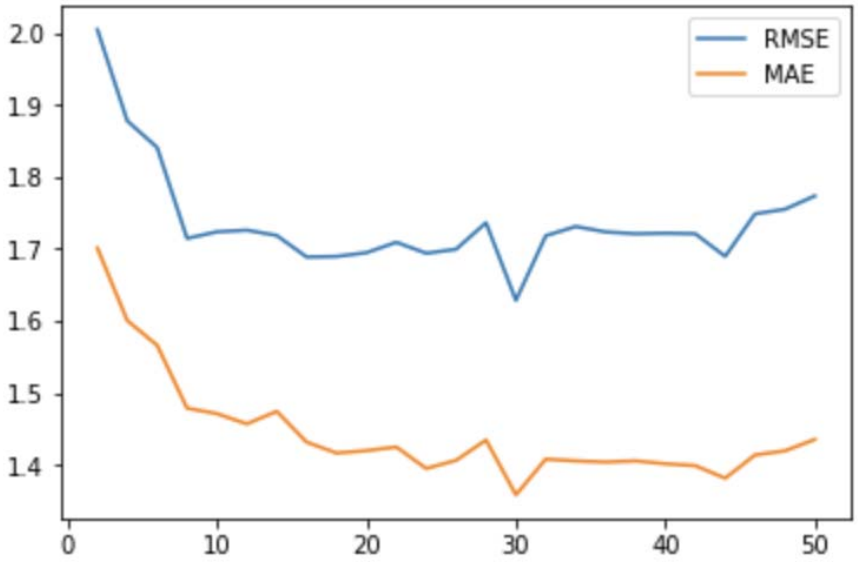
Question 20 is the same as Question 19 except the trimming part. The only difference between them is that we picked unpopular movies, whose rating number is less than 2, as testset at this time. The result is shown in the figure. The minimum average RMSE is 1.231 when k is 48. It is obvious that for unpopular movies, their average RMSE and MAE values mainly decrease monotonically.



RMSE/MAE Plot

Question 21

Similar to the former questions, we change the trimming function and picked the movies that have variance of at least 2 and more than 5 ratings in Question 21. The RMSE/MAE plot is shown in the following figure. The minimum average RMSE is 1.629 when k is 30. Compared with the plot in Question 20, RMSE/MAE values fluctuate more strongly this time. But in general, RMSE/MAE still decreases as k increases.



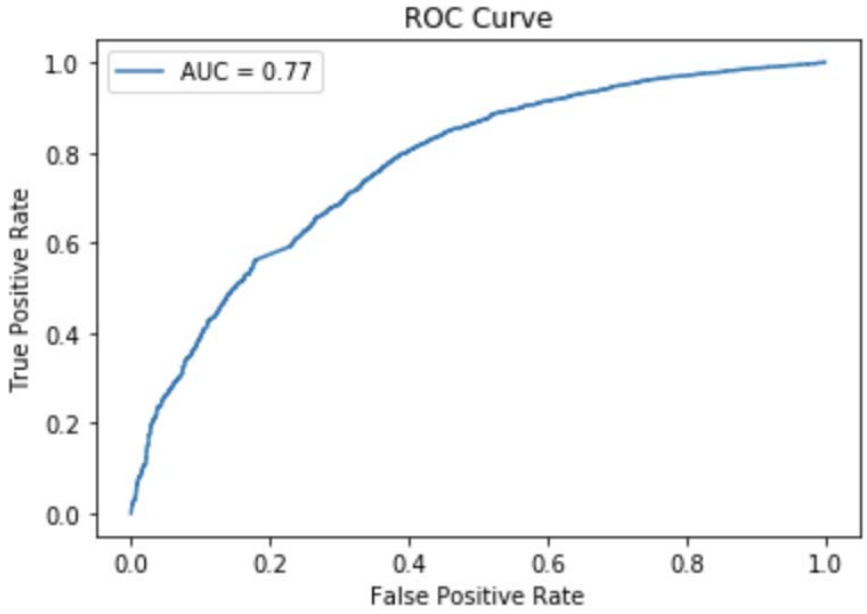
RMSE/MAE Plot

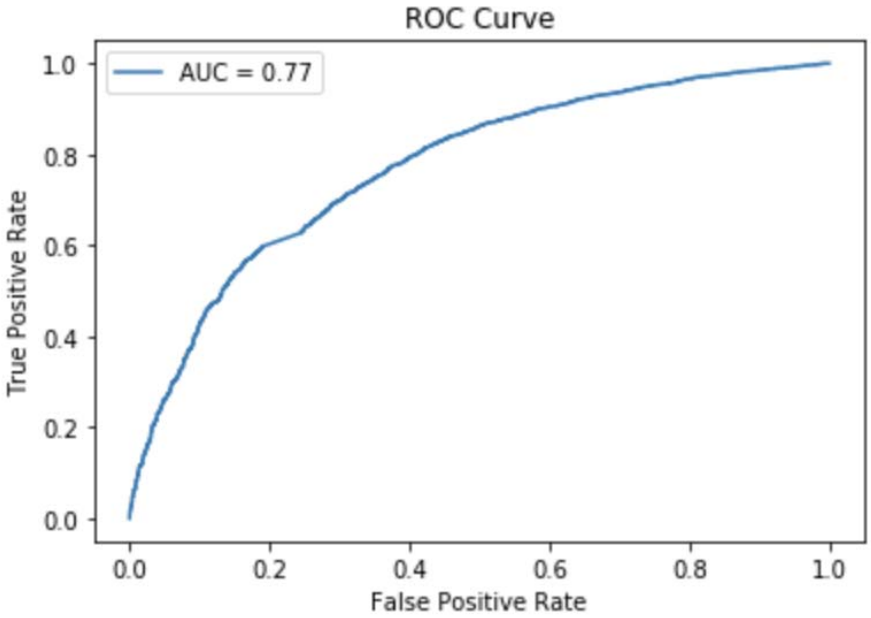
Question 22

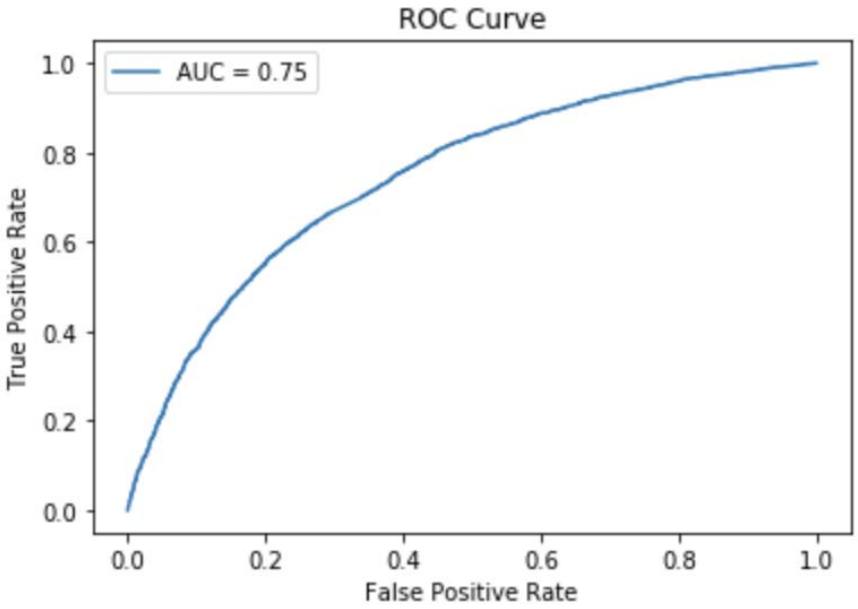
In this question, we set k as 20 and swept threshold value through the list [2.5, 3, 3.5, 4]. The ROC plots we got are shown in the following figures. Their corresponding AUC values are listed in the table.

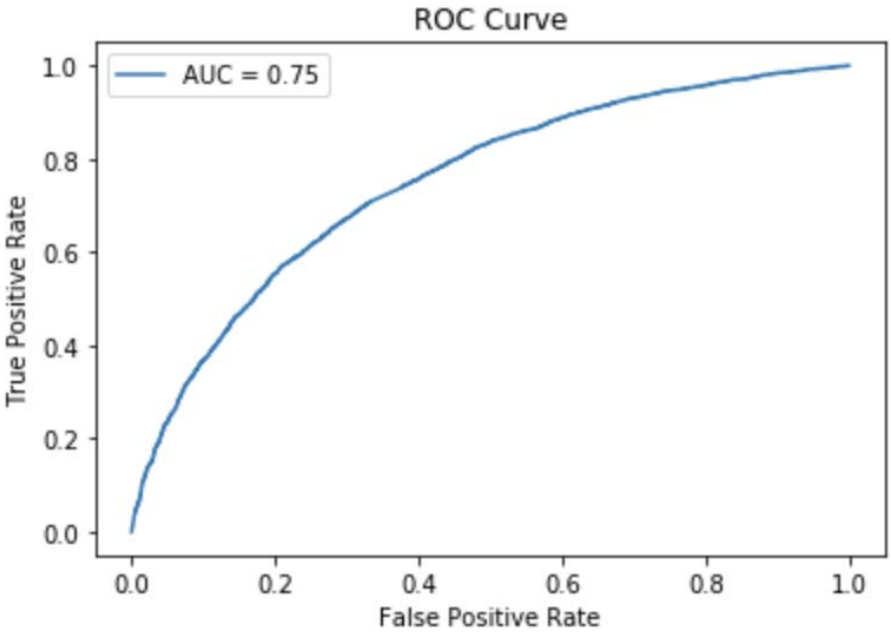
AUC Value Table

Threshold	2.5	3	3.5	4
AUC	0.77	0.77	0.75	0.75




ROC Curve
AUC = 0.77
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0


ROC Curve
AUC = 0.75
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0



Question 23

We selected the first 5 columns in the V matrix and listed the genres of the top 10 movies.

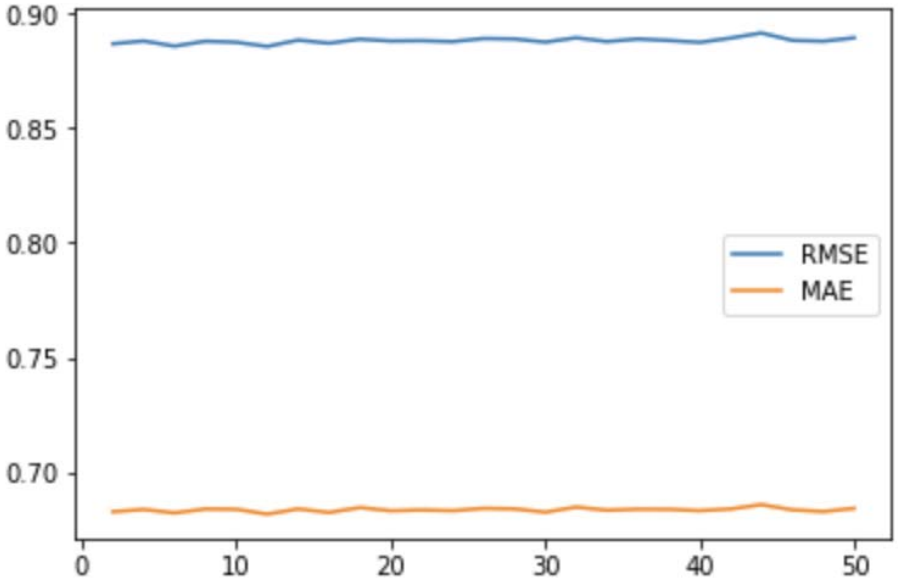
Column 1	Drama | War	Drama | Mystery | Romance	Action | Comedy | Crime | Fantasy	Action | Adventure | Sci-Fi | War | IMAX	Comedy | Documentary
	Children | Comedy	Drama	Comedy | Romance	Comedy	Action | Adventure | Animation | Children | Comedy | Romance
Column 2	Adventure | Animation	Comedy | Documentary	Musical	Drama | Fantasy | Horror	Comedy
	Comedy | Romance	Drama | War	Drama | Fantasy | Mystery | Romance	Drama | Romance | War	Comedy
Column 3	Adventure | Drama | Fantasy | Romance	Action | War	Action | Adventure | Drama | Thriller	Comedy	Action | Crime | Thriller

	Comedy | Drama | Romance	Comedy	Comedy	Comedy | Mystery | Thriller	Comedy | Drama
Column 4	Drama	Documentary | Drama	Drama | Romance	Drama | Thriller	Comedy | Crime
	Comedy | Drama	Adventure | Children	Action	Documentary	Comedy
Column 5	Comedy | Crime | Mystery | Thriller	Children | Comedy | Musical | Romance	Comedy	Drama | Romance	Action | Adventure | Sci-Fi | Thriller
	Documentary	Comedy | Romance	Documentary	Drama | Musical | Romance	Adventure | Animation | Children | Musical | Western

We note that there are 18 genres totally in the table. As we said in the Question 18, the optimal number of latent factors is very close to or even the same as the movie genres. Meanwhile, the genres of the top 10 movies generally concentrate in the collection of Comedy and Drama.

Question 24

In this question, we changed the NMF filter to a MF with bias collaborative filter and repeated the same prediction and plotting processes. To achieve such a MF filter, we used the SVD model in surprise package. The average RMSE and MAE lines are plotted in the following figure.



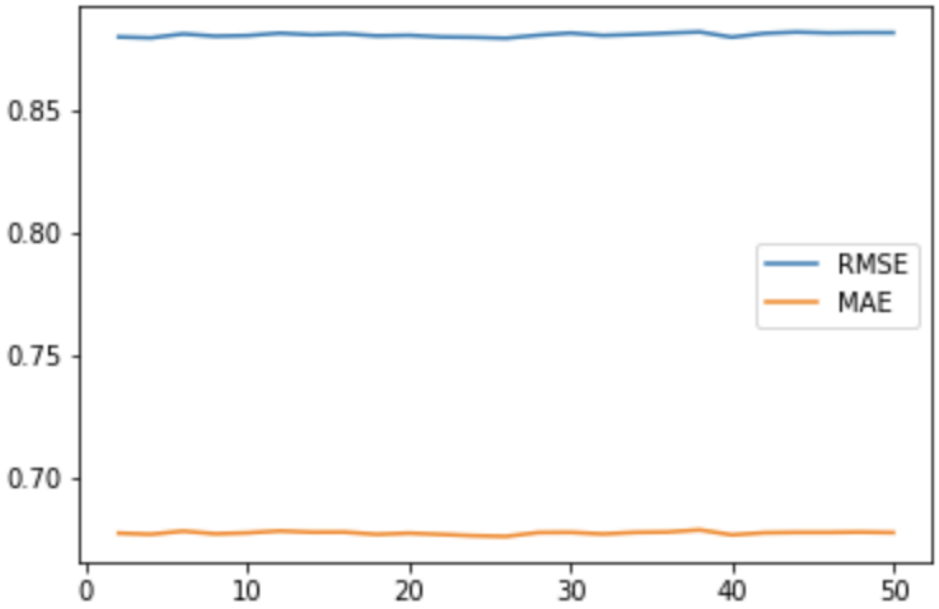
RMSE/MAE Plot

Question 25

We can note that average RMSE/MAE values almost keep constant as k changes. According to the plot, the optimal number of latent factors is 12 for both RMSE and MAE. The average RMSE can reach its minimum value, 0.886 when k = 20, and the minimum of average MAE is 0.682.

Question 26

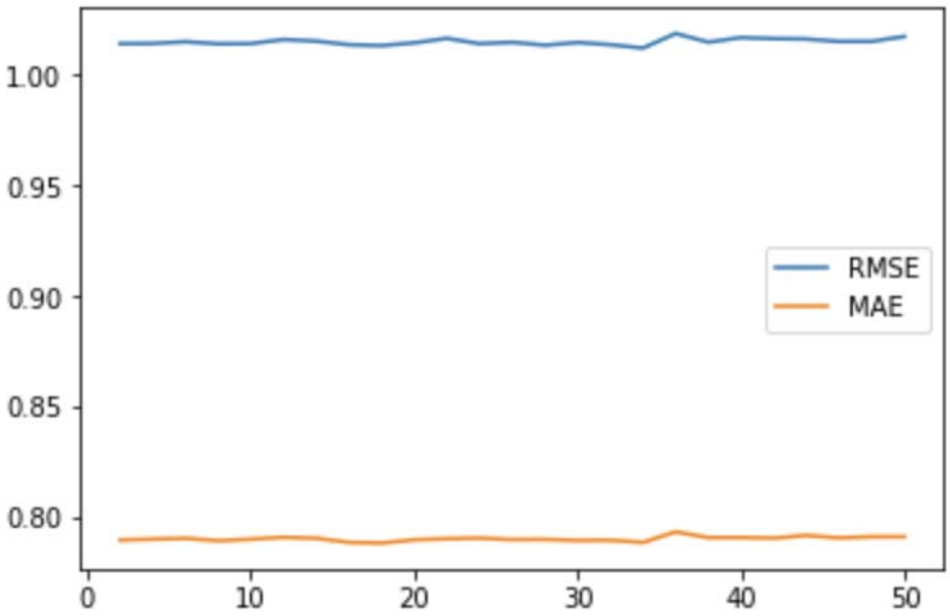
Just like the former questions, we need to complete the popular trimming, unpopular trimming and high-variance trimming on the testset in Question 26, 27, 28 respectively. For this part, we trimed the testset data before prediction and repeat the above steps. The values of average RMSE/MAE against k are plotted in the following figure. The minimum average RMSE is 0.879 when k is 26.



RMSE/MAE Plot

Question 27

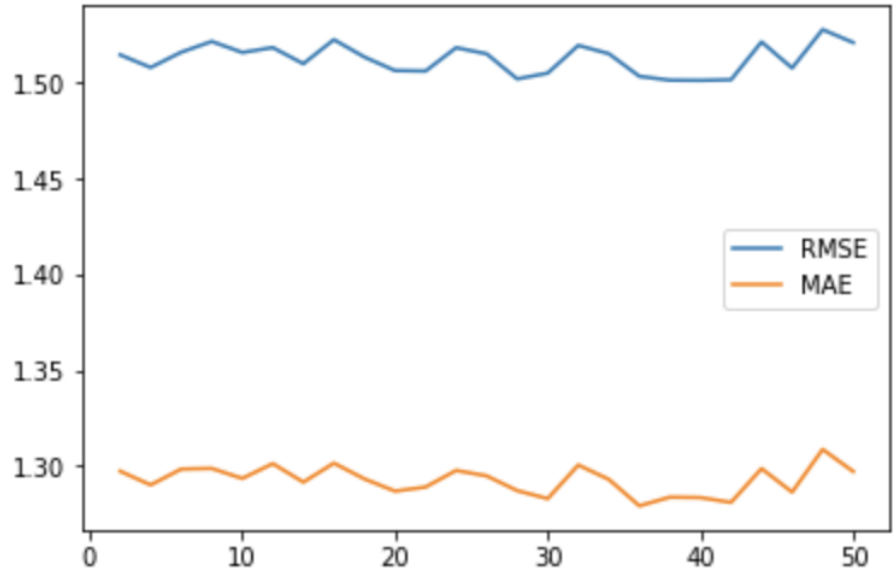
With the unpopular movie trimmed test set, we got the RMSE/MAE vs k plot as follow. The minimum average RMSE is 1.012 when k = 34.



RMSE/MAE Plot

Question 28

For high variance trimming, we have the following RMSE/MAE curves and the minimum average RMSE is 1.501 as k is 40.



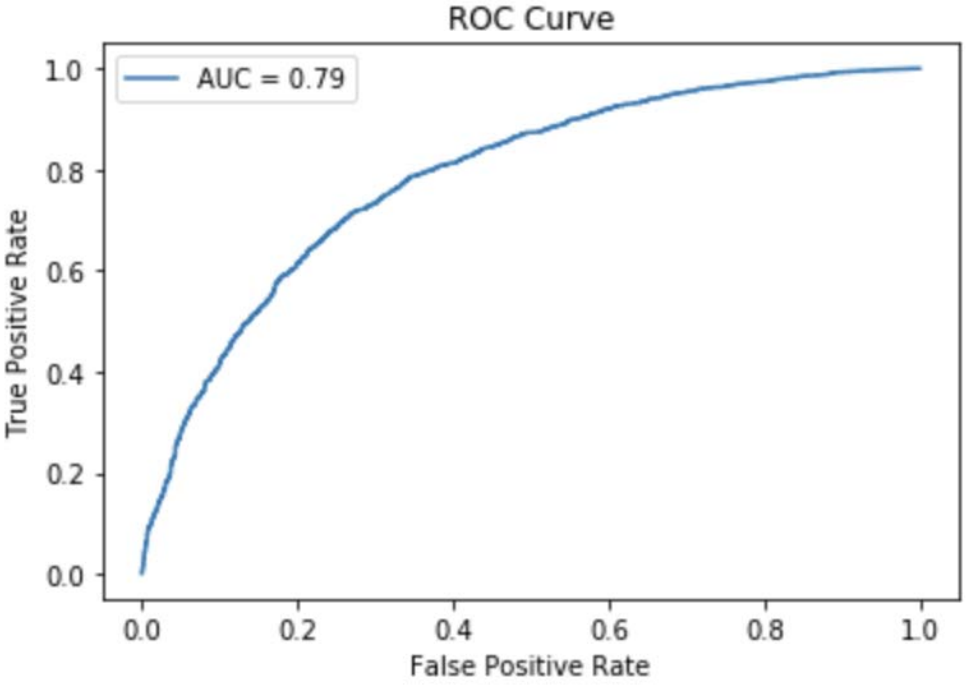
RMSE/MAE Plot

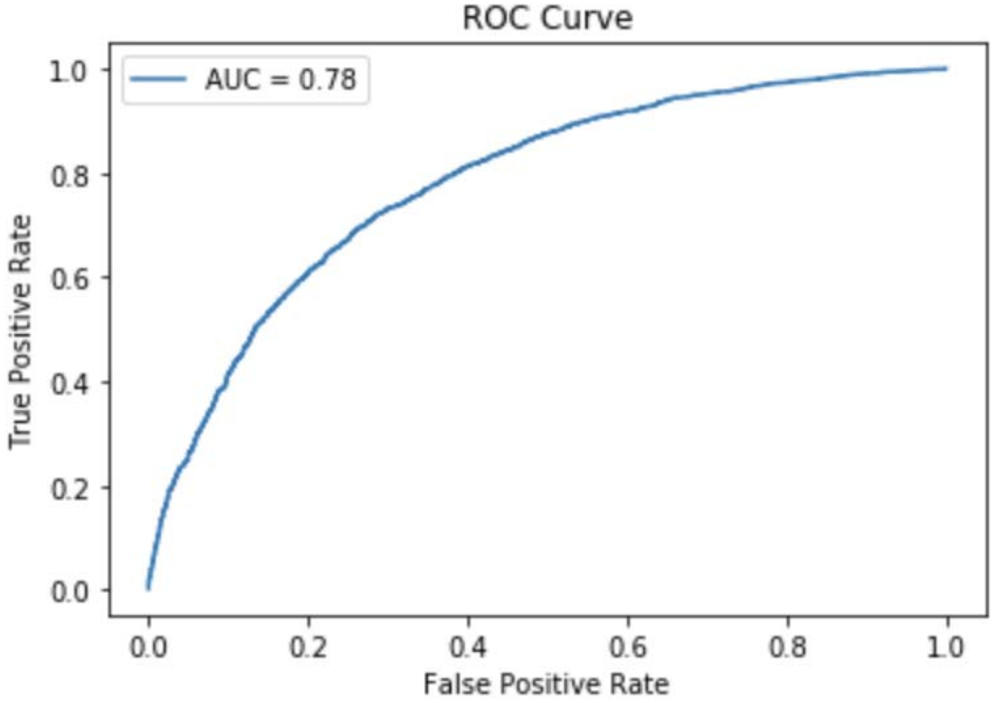
Question 29

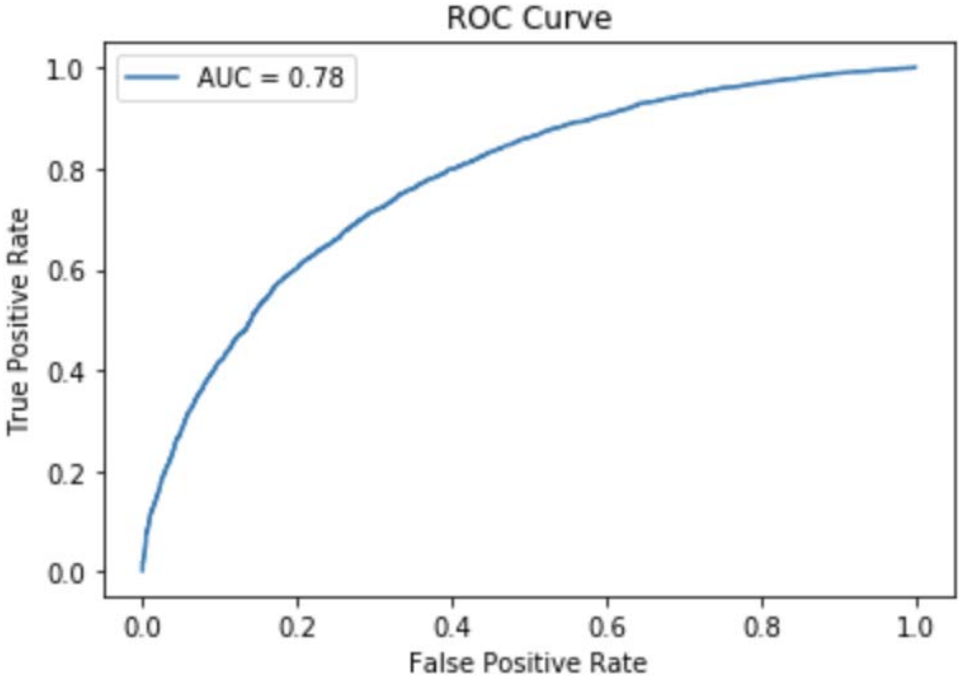
In this question, we set k as 12 and swept threshold value through the list [2.5, 3, 3.5, 4]. The ROC plots we got are shown in the following figures. Their corresponding AUC values are listed in the table.

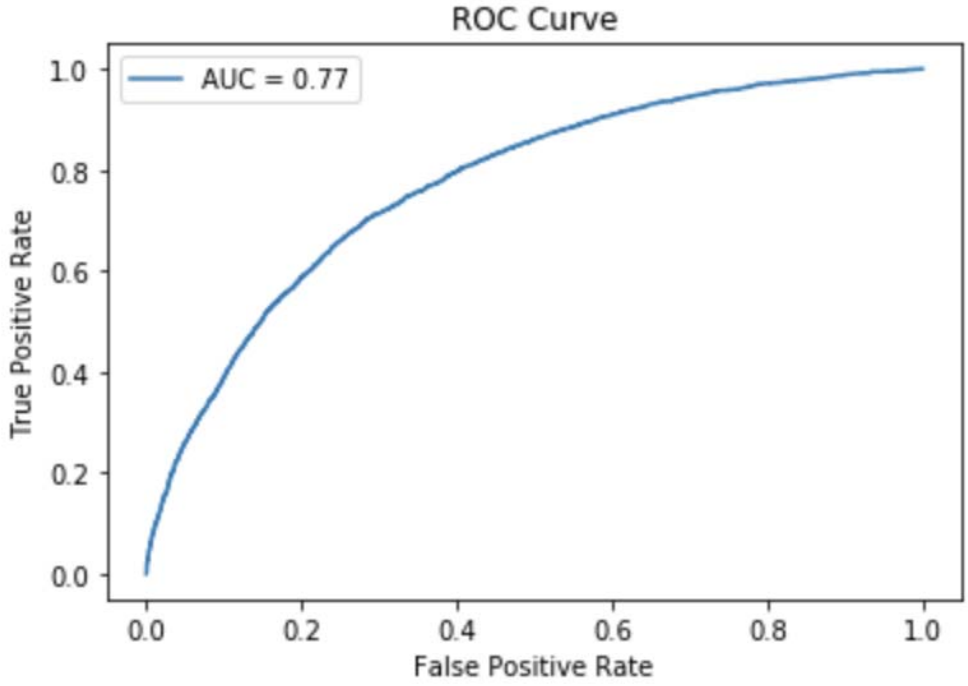
AUC Value Table

Threshold	2.5	3	3.5	4
AUC	0.79	0.78	0.78	0.77


ROC Curve
AUC = 0.79
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0


ROC Curve
AUC = 0.78
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0





6. Naïve Collaborative Filtering

Question 30

Average RMSE using Naive collaborative filtering prediction is 0.9104001601499685.

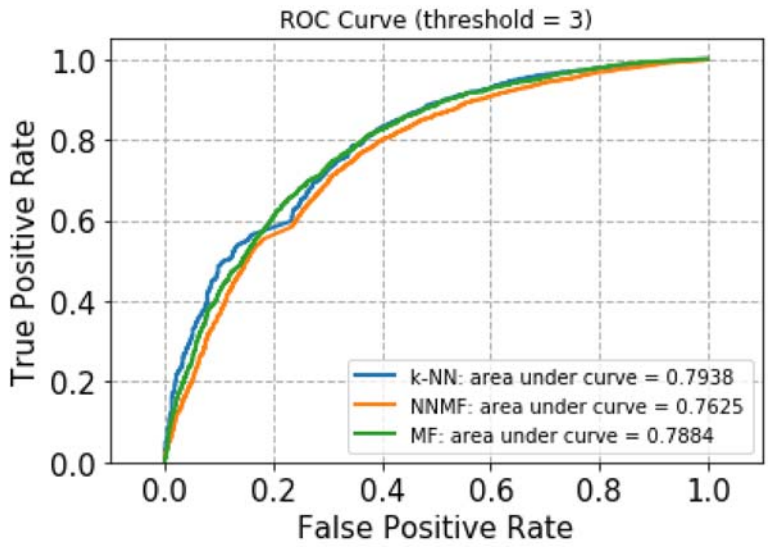
Question 31
Average RMSE using Naive collaborative filtering prediction in popular set is 0.9064083550344585.

Question 32
Average RMSE using Naive collaborative filtering prediction in unpopular set is 0.6537427799366957.

Question 33
Average RMSE using Naive collaborative filtering prediction in high variance set is 0.7447896222756207.

7. Performance Comparison

Question 34



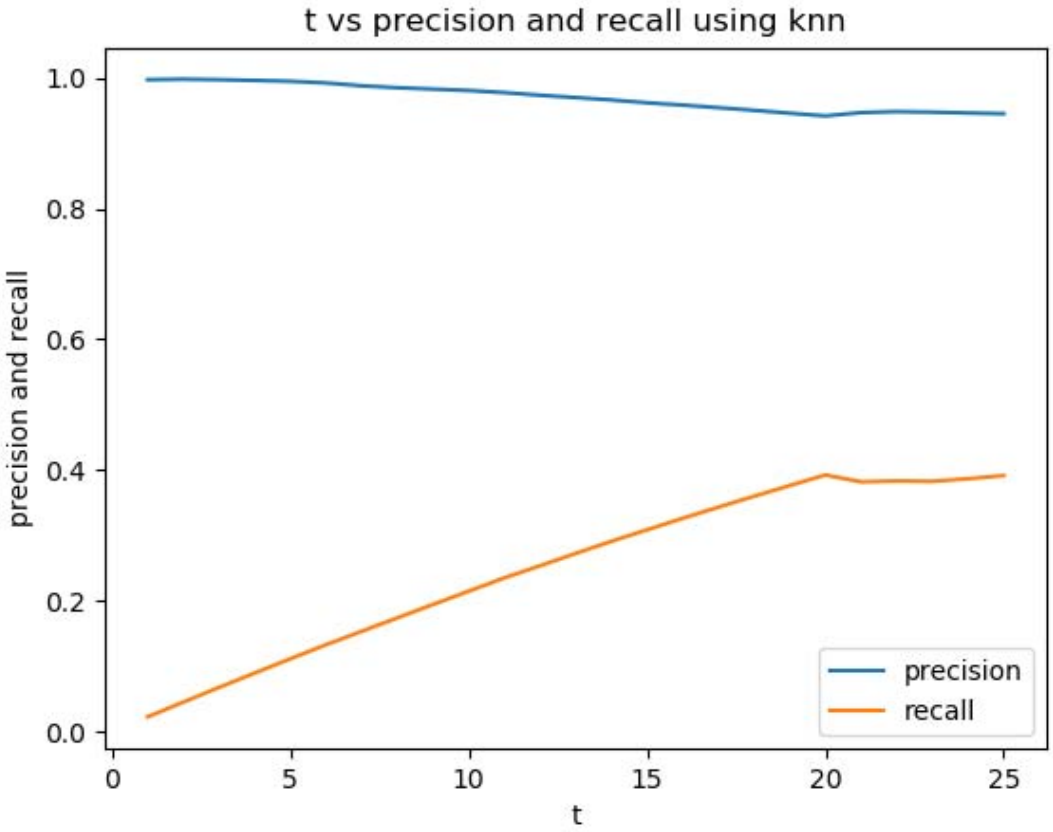
In this question we plot the ROC curves for k-NN, NNMF, and MF with bias based collaborative filters in the same figure, using threshold value = 3. To have a better comparison, we set the parameter random_state = 0 in the train_test_split function so that the three filters predicted on a same testset. From the above plot we can see that there's only slight difference between these three curves. If we look at the AUC value, we can say that k-NN collaborative filter has the best prediction result with the highest AUC value = 0.7938, and then the NNMF collaborative filter, with an AUC value of 0.7884. MF with bias based collaborative filter's performance is not as good as the other two, with only an AUC value of 0.7625.

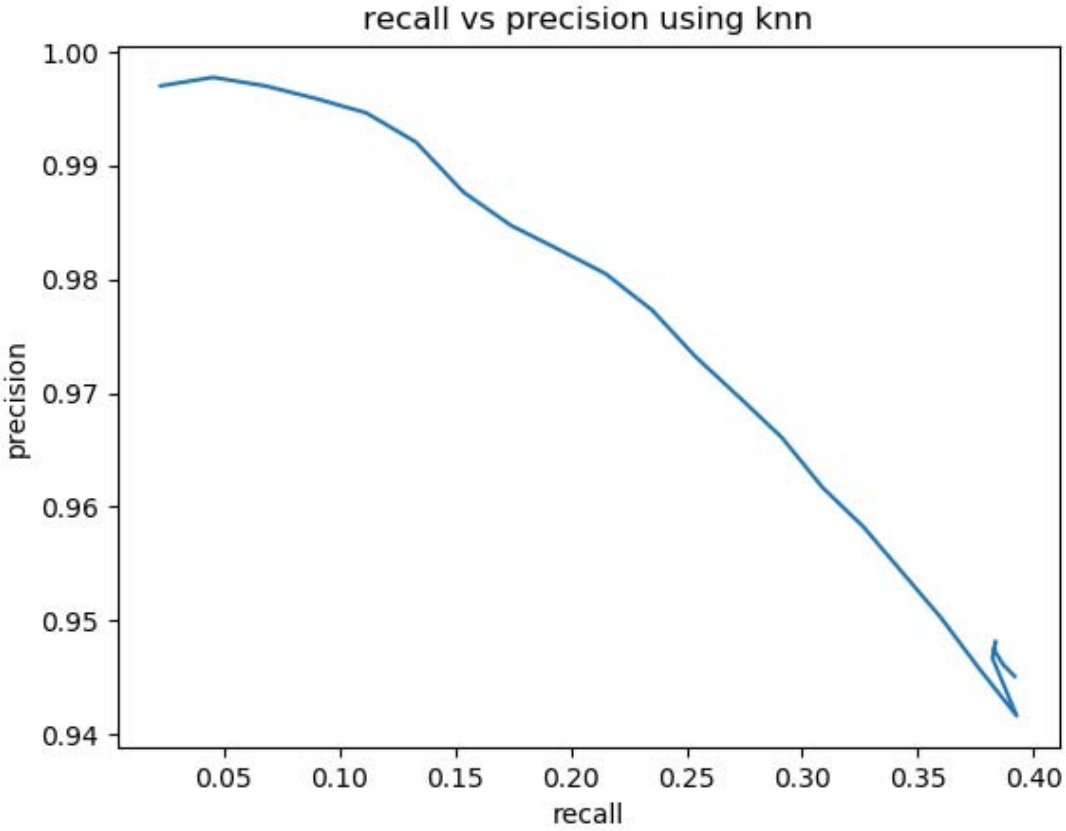
8. Ranking

Question 35

Precision: in the movies we recommended to the user, the percent of movies that the user actually likes.
Recall: in the movies the user likes, the percent of movies that we did recommend to him.

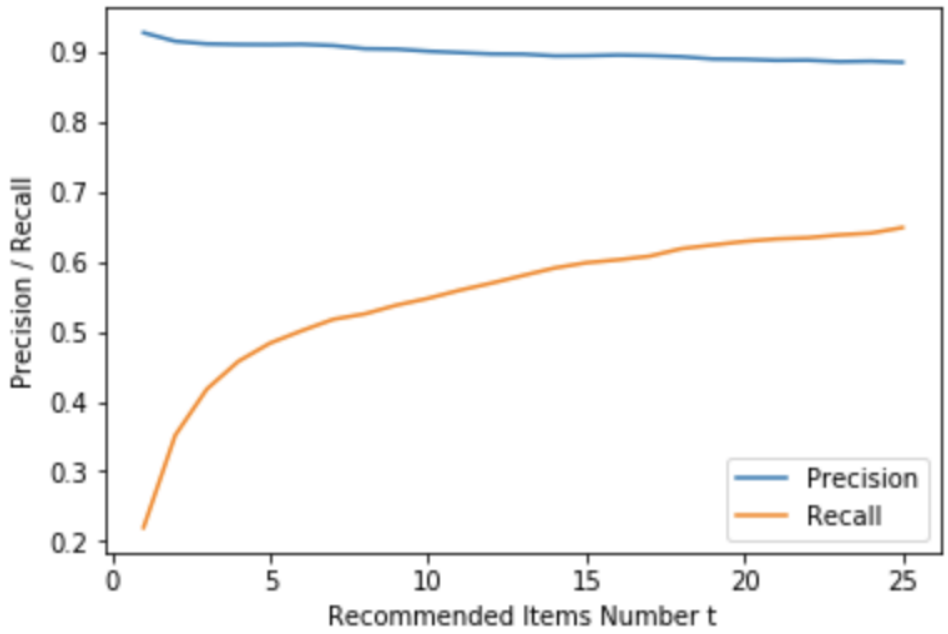
Question 36



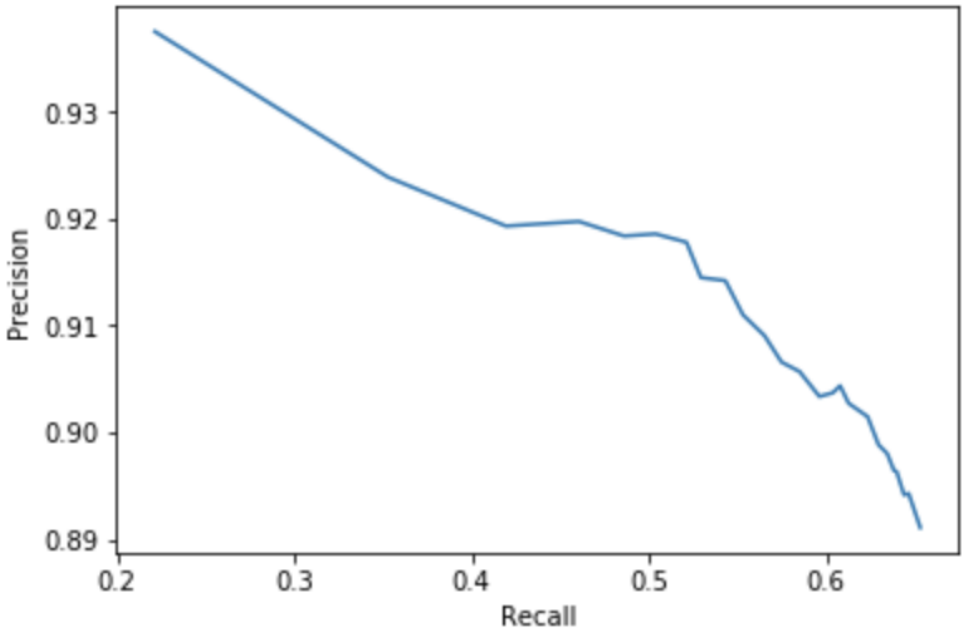


As t grows, the precision keeps falling slowly, and recall keeps growing steadily in a linear manner. And at t=20, they both seem to reach a plateau and change less than before, become quite stable.
As recall grows, precision keeps falling.

Question 37

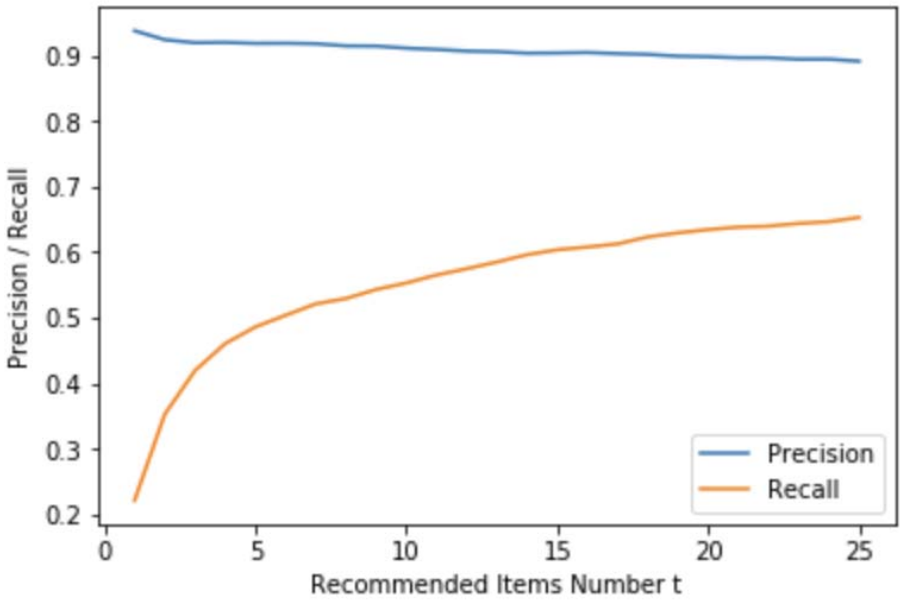


As the recommended item number t increases, the precision will decrease very slowly so that the precision is always around 0.9. Meanwhile, the recall will increase fast and end up at about 0.65.

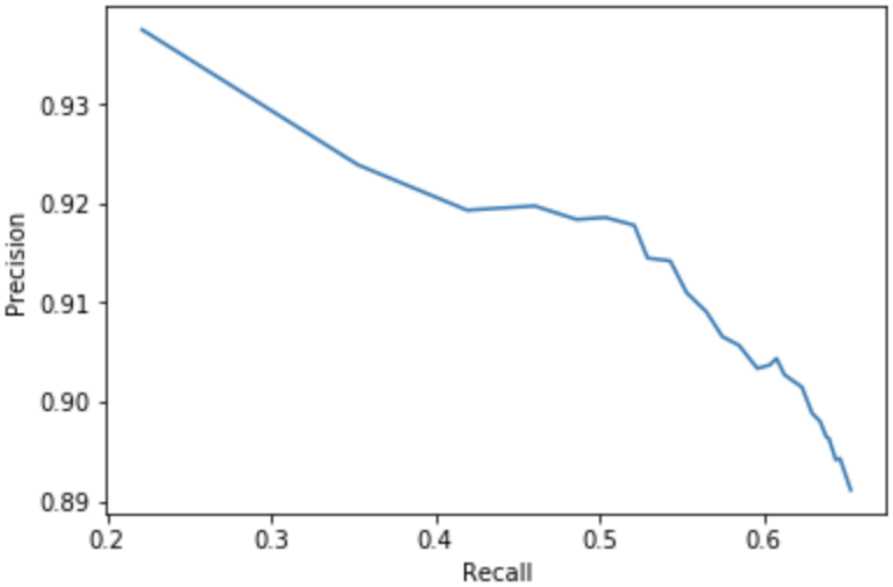


As recall increases, the precision almost decreases in a linear way.

Question 38

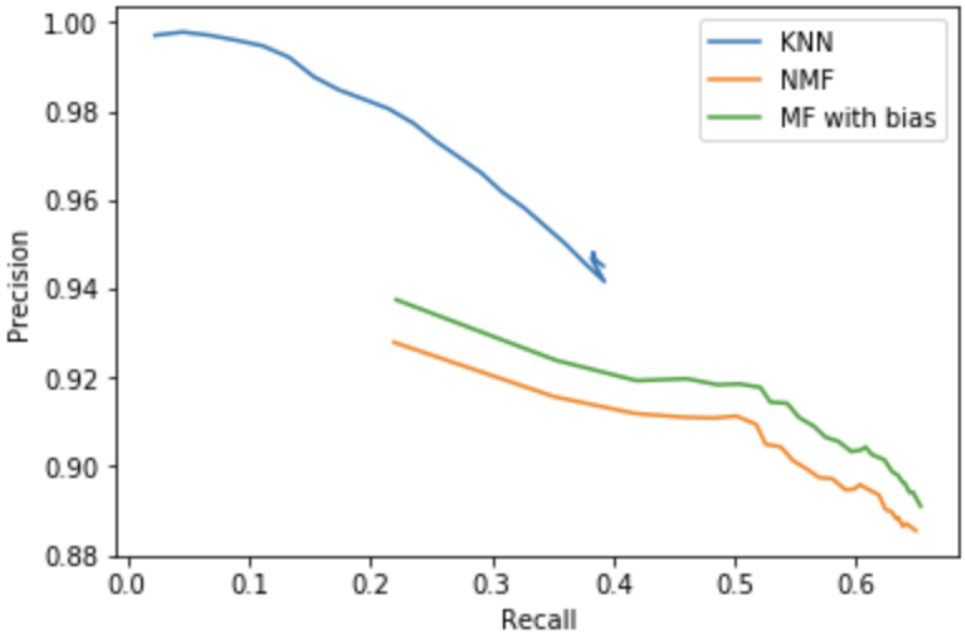


The MF with bias filter case is very similar to that of NMF. As the recommended item number t increases, the precision will decrease very slowly so that the precision is always around 0.9. Meanwhile, the recall will increase fast and end up at about 0.65.



As recall increases, the precision almost decreases in a linear way.

Question 39



According to the above plot, we can find that the KNN prediction enjoys the best precision, but the worst recall. The case in NNMF is very similar to that of MF with bias. When they have the same recall value, MF with bias filter has a better precision.