Project 3: Collaborative Filtering

Name: Jianxiong Wang, Yijun Wu, Yanzhao Wang, Yutong Sun

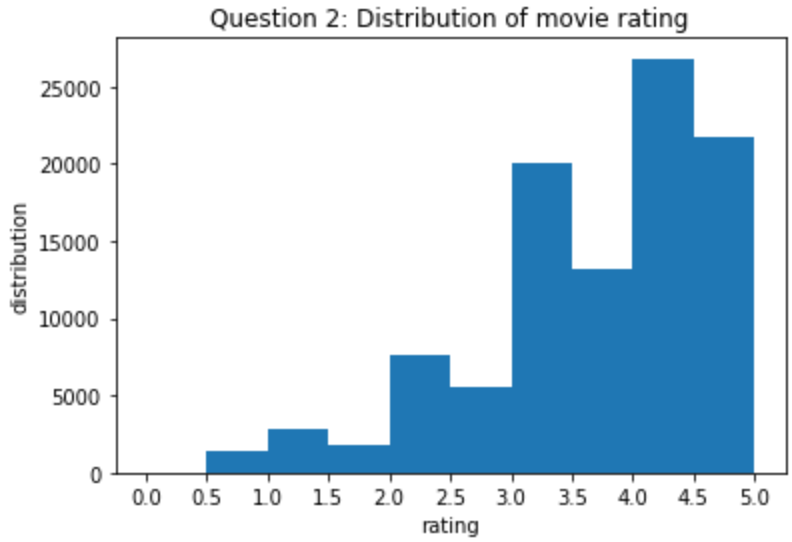
Date: 2019.2.14

Question 1

Compute the sparsity of the movie rating dataset, where sparsity is defined by equation 1
Total number of possible ratings: 5942620
Total number of available ratings: 100836
Sparsity 0.016968273253211548

Question 2

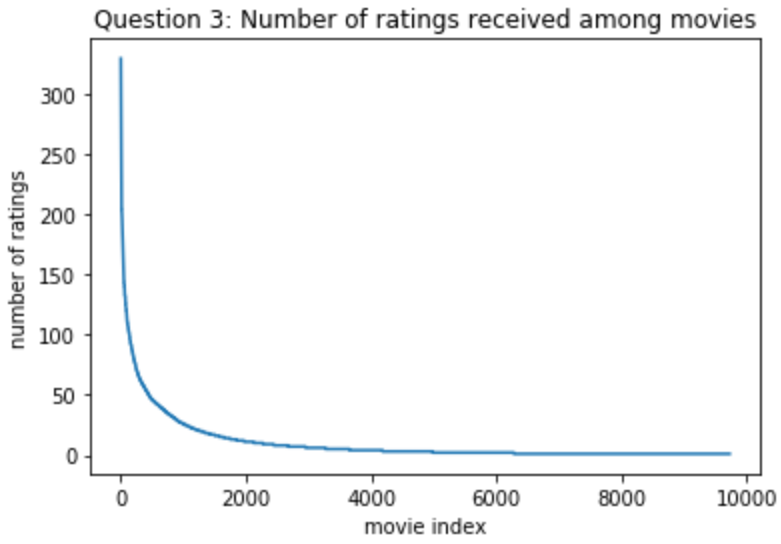
Plot a histogram showing the frequency of the rating values. To be specific, bin the rating values into intervals of width 0.5 and use the binned rating values as the horizontal axis. Count the number of entries in the ratings matrix R with rating values in the binned intervals and use this count as the vertical axis. Briefly comment on the shape of the histogram



The distribution is left-skewed. People are more willing to rat the movies they like than rating the movies they dislike.

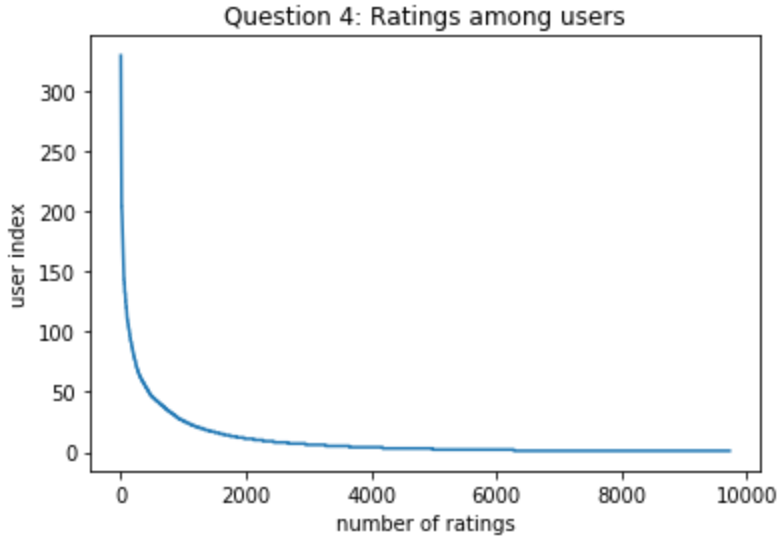
Question 3

Plot the distribution of the number of ratings received among movies. To be specific, the X-axis should be the movie index ordered by decreasing frequency and the Y -axis should be the number of ratings the movie has received. For example, the movie that has the largest number of ratings has index 1; ties can broken in any way. A monotonically decreasing curve instead of a histogram is expected.



Question 4

Plot the distribution of ratings among users. To be specific, the 2 X-axis should be the user index ordered by decreasing frequency and the Y -axis should be the number of movies the user have rated. The requirement of the plot is similar to that in Question 3.



Question 5

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

The distribution has a really long tail, and this means that only a small portion of the movies have lots of ratings most of the movies have a small number of ratings. This means that the matrix is very sparse. The sparse matrix limits the coverage of neighborhood-based methods, and this creates challenges for robust similarity computation. But popular movie is easier to predict.

Question 6

Compute the variance of the rating values received by each movie. Then, bin the variance values into intervals of width 0.5 and use the binned variance values as the horizontal axis. Count the number of movies with variance values in the binned intervals and use this count as the vertical axis. Briefly comment on the shape of the histogram.

The distribution means that the ratings on the movies are very consistent. Most of the movie have very low variance, and only small amount of movie have high variance in terms of reviews. Therefore, it is easier to predict popular moives and harder to predict not so popular movies.

Question 7

Write down the formula for μ u in terms of I u and r uk

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

Question 8

In plain words, explain the meaning of I u ∩ I v . Can I u ∩ I v = ⌀ (Hint: Rating matrix R is sparse).
This means the set of items have been rated both by user u and user v. When the two sets are disjoint these two sets have not rated any common item.

Question 9

Can you explain the reason behind mean-centering the raw ratings in the prediction function? (Hint: Consider users who either rate all items highly or rate all items poorly and the impact of these users on the prediction function)

Different users may have different baselines for rating movies. Some users tend to give higher scores for all movies while some do the opposite. If a movie is rated by mostly “lenient” users, it would have a better score than if it is viewed by the average users. This uncertainty is not wanted because the movie’s rating changes simply as a result of which users have rated it, which is not relevant to the movie itself.

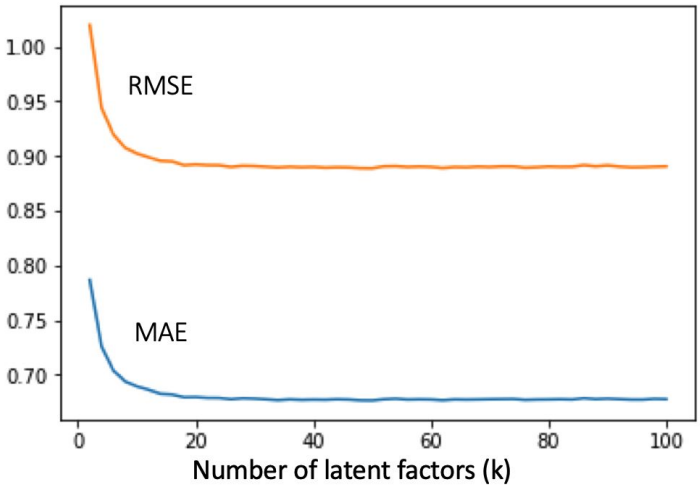
In order to solve the problem, we subtract the average rating from the user so that the user bias is eliminated. The rating now would only contain preferences about the movies.

Question 10, 11

Design a k-NN collaborative lter to predict the ratings of the movies in the MovieLens dataset and evaluate it's performance using 10-fold cross validation. Sweep k (number of neighbors) from 2 to 100 in step sizes of 2, and for each k compute the average RMSE and average MAE obtained by averaging the RMSE and MAE across all 10 folds. Plot average RMSE (Y-axis) against k (X-axis) and average MAE (Y-axis) against k (X-axis).

Use the plot from question 10, to and a 'minimum k'. Note: The term 'minimum k' in this context means that increasing k above the minimum value would not result in a signicant decrease in average RMSE or average MAE. If you get the plot correct, then 'minimum k' would correspond 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

Result:



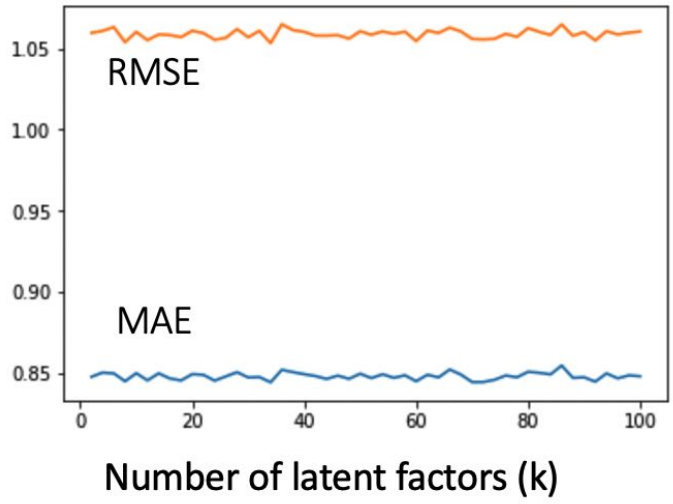
The minimum MAE is 0.6766.
The minimum RMSE is 0.8887.

As we can see from the figure, both error converges to a constant at about 18. That is the “minimum k” that we would use in later questions.

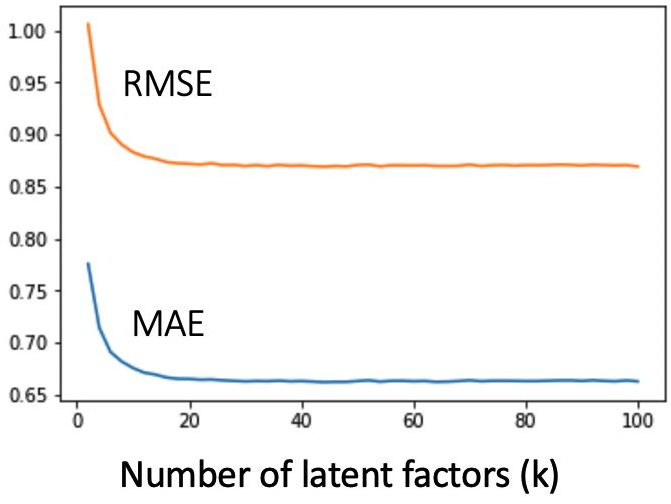
Question 12-14

Design a k-NN collaborative filter to predict the ratings of the movies in the popular, unpopular, high-variance movie trimmed test set and evaluate it's performance using 10-fold cross validation. Sweep k (number of neighbors) from 2 to 100 in step sizes of 2, and for each k compute the average RMSE obtained by averaging the RMSE across all 10 folds. Plot average RMSE (Y-axis) against k (X-axis). Also, report the minimum average RMSE

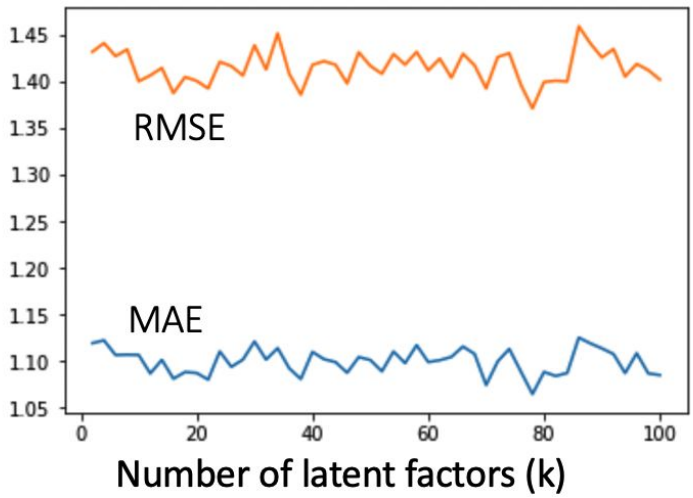
```
Unpopular Movies
Minimum MAE =  0.8441870772549814
Minimum RMSE =  1.0533907422876287
```




```
Popular Movies
Minimum MAE =  0.6619676431071577
Minimum RMSE =  0.8689759667922179
```



```
High-variance Movies
Minimum MAE =  1.0648513526545365
Minimum RMSE =  1.371434304693082
```

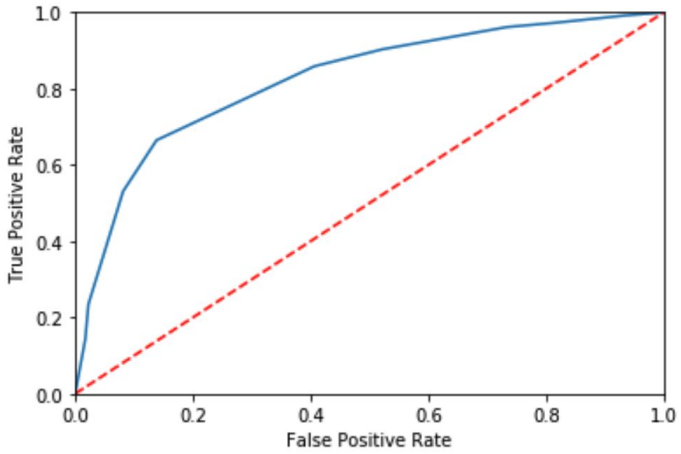


As we can see from the figures, the performance of k-NN on popular movies is similar to and slightly better than on the whole data set. The similarity in performance is expected because over 90% of all movies fall into the category of “popular movies”. The performance on unpopular movies are quite stable. Both errors are significantly higher than the ones of popular movies. The performance on high-variance movies are unstable, RMSE ranging from 1.35 to 1.45, MAE ranging from 1.07 to 1.13. The errors on high-variance movies are much higher than the ones on popular and unpopular movies. (1.4 vs 1.05)

Question 15

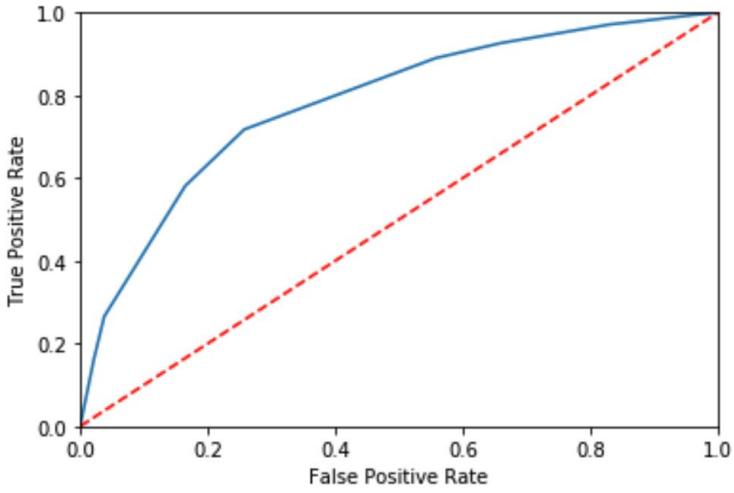
Plot the ROC curves for the k-NN collaborative Iter designed in question 10 for threshold values [2:5; 3; 3:5; 4]. For the ROC plotting use the k found in question 11. For each of the plots, also report the area under the curve (AUC) value.

Threshold = 2.5



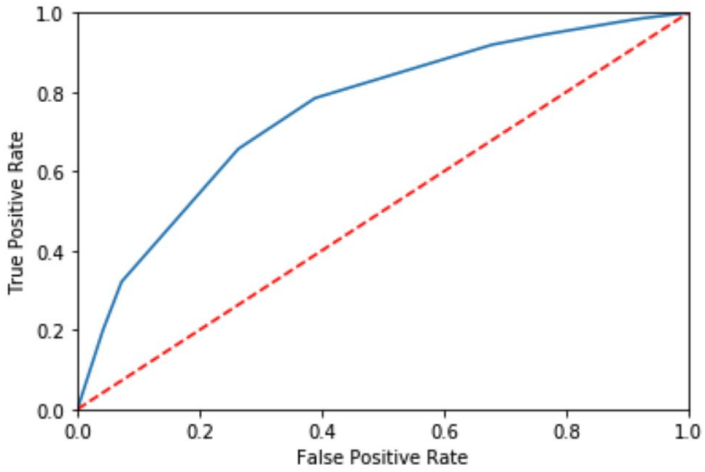

```
AUC =  0.8234930719945448
```

Threshold = 3.0



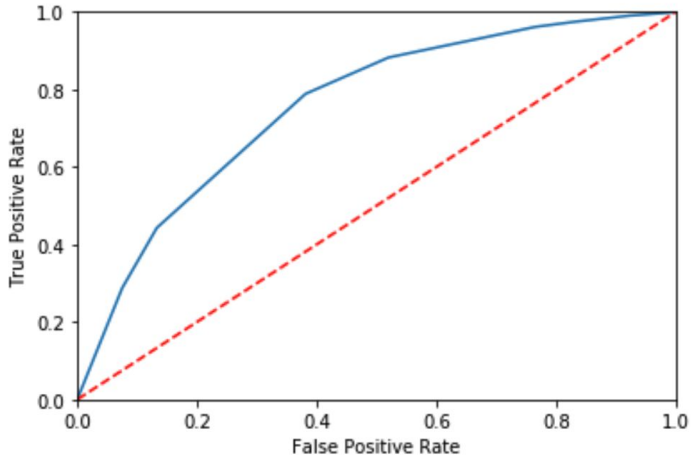

```
AUC =  0.7824723009062248
```

Threshold = 3.5




```
AUC =  0.7525169669665743
```

Threshold = 4.0




```
AUC =  0.7573877077501963
```

Question 16

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.

The optimization problem is not convex as its Hessian matrix is not positive definite. However, if U is fixed, this becomes a convex optimization problem. It becomes a least-squares problem with regulation on V.

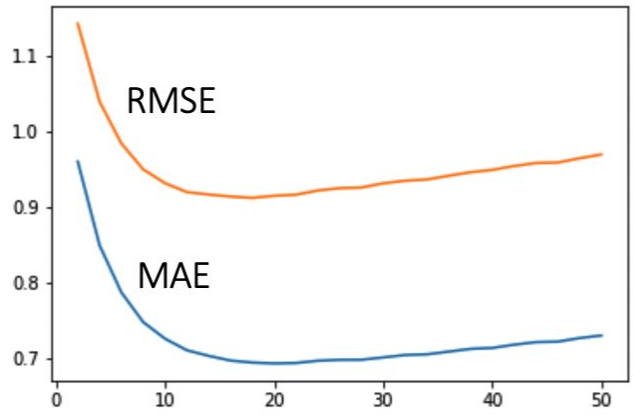
Minimize over V $\sum_{i=1}^{m}\sum_{j=1}^{n} W_{ij}(r_{ij} - (UV^T)_{ij})^2 + \lambda|V|^2$

Subject to $V \geq 0$

Question 17-18

Design a NNMF-based collaborative Iter to predict the ratings of the movies in the MovieLens dataset and evaluate it's performance using 10-fold cross-validation. Sweep k (number of latent factors) from 2 to 50 in step sizes of 2, and for each k compute the average RMSE and average MAE obtained by averaging the RMSE and MAE across all 10 folds. 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. Use the plot from question 17, to nd 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?

```
Best k for MAE is:  20 MAE =  0.6934550041790116
Best k for RMSE is:  18 RMSE =  0.9121200067403998
```



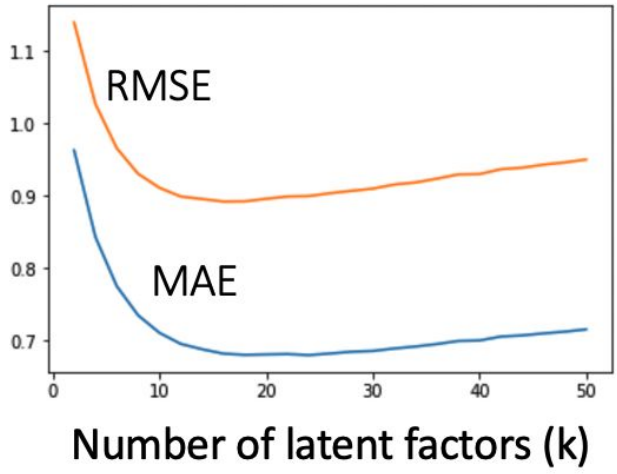
The number of genres is 20. The optimal number of latent factors for MAE is the same. However the optimal number for RMSE is not, but very close.

Question 18-20

Use the plot from question 17, to nd 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?

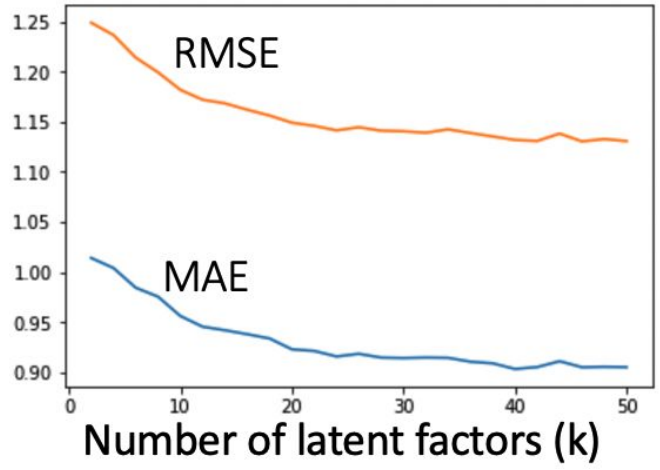
Popular Movies

```
Popular Movies
Best k for MAE is:  24 MAE =  0.6803685954443075
Best k for RMSE is:  16 RMSE =  0.8925160454931088
```



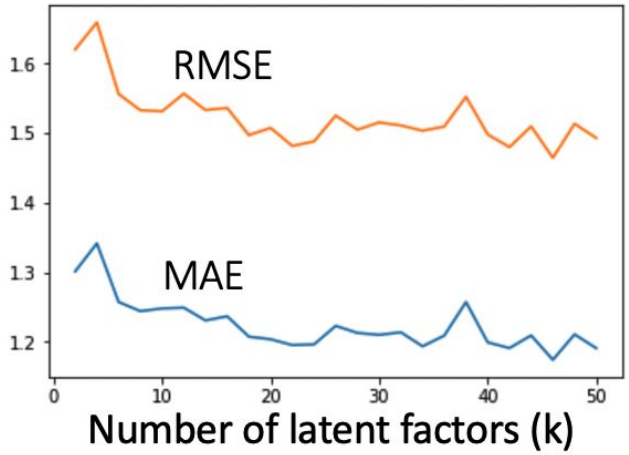
Unpopular Movies

```
Unpopular Movies
Best k for MAE is:  40 MAE =  0.9036574125514127
Best k for RMSE is:  46 RMSE =  1.130490373196088
```



High-Variance Movies

```
High-variance Movies
Best k for MAE is:  46 MAE =  1.1741147455766598
Best k for RMSE is:  46 RMSE =  1.4645825387093534
```

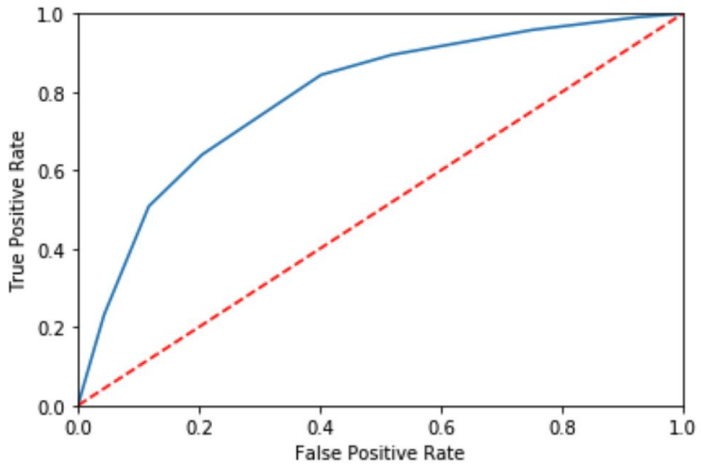


The optimal number of latent factors for the three models and two error measures are different from the number of movie genres. It is worth mentioning that the ones for popular movies are quite close(16/24 vs. 20) and the trend of both errors over k is similar. However the optimal number of latent factors for unpopular and high-variance movies are very different and the trend of the errors are different, containing no apparent uptick over k larger than 20.

Question 22

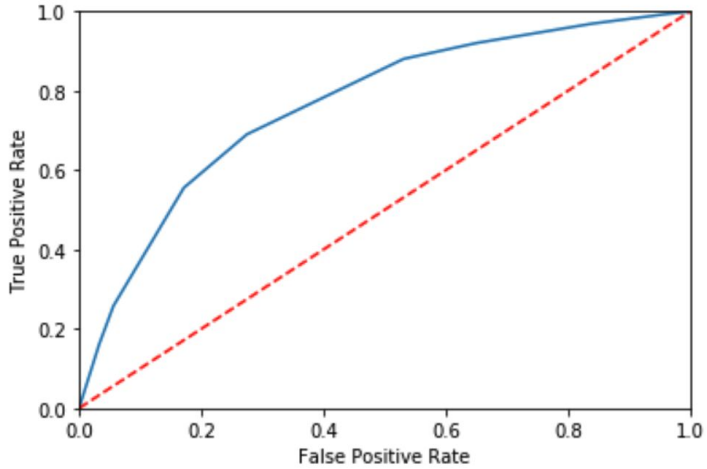
Plot the ROC curves for the NNMF-based collaborative filter designed in question 17 for threshold values [2:5; 3; 3:5; 4]. For the ROC plot-ting use the optimal number of latent factors found in question 18. For each of the plots, also report the area under the curve (AUC) value.

Threshold = 2.5



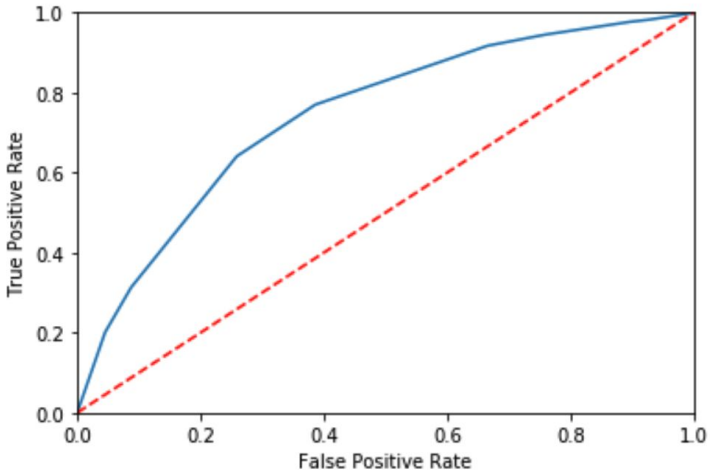

```
AUC =  0.7899146003655058
```

Threshold = 3.0



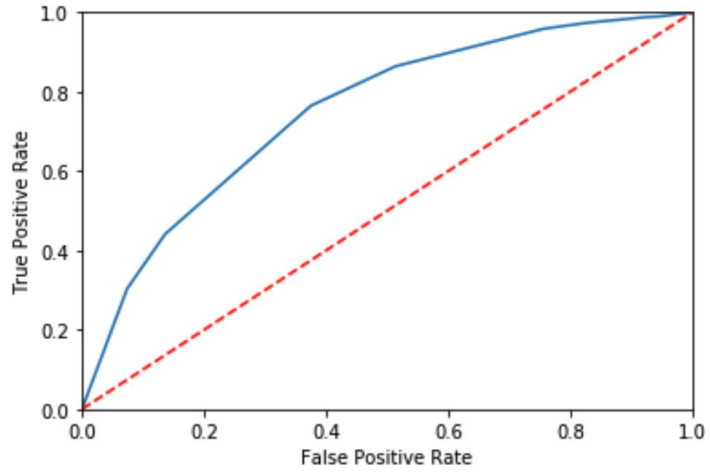

```
AUC =  0.7642492383327988
```

Threshold = 3.5




```
AUC =  0.7438724450034335
```

Threshold = 4.0




```
AUC =  0.7512962764389263
```

Question 23

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?

The top 10 movies of each latent factor have a much higher similarity in their genres. For example, this is the genres of top 10 movies of latent factor 7. Note that the majority (6) of the movies here are comedy.

```
i =  6
Comedy
Horror|Sci-Fi
Drama
Drama|Musical|Romance
Comedy|Crime|Thriller
Comedy
Horror|Sci-Fi
Comedy|Crime|Drama|Thriller
Comedy
Comedy|Drama|Romance
```

The half of top 10 movies in this latent factor is Thriller.

```
i =  1
Horror
Comedy|Drama|Fantasy|Romance
Crime|Drama|Musical
Mystery|Thriller
Action|Drama|Thriller
Action|Crime|Drama|Thriller
Drama|Thriller
Drama|Thriller
Drama
Children|Drama|Fantasy
```

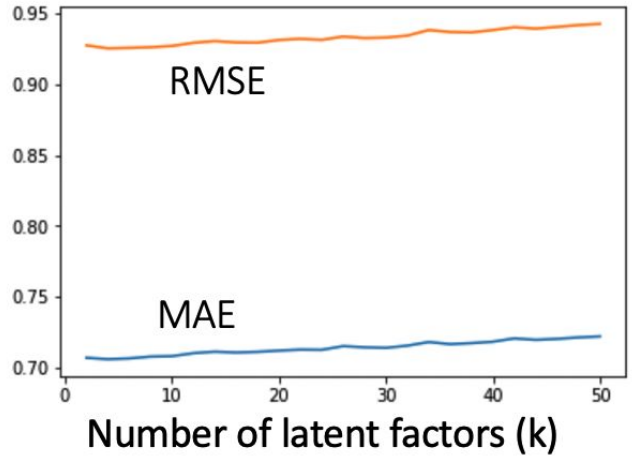
As we can see from these two columns, the top 10 movies of each latent factor are often from few genres instead of being mixed. This shows the effect of NNMF to distinguish different styles of movies which results in the concentration of certain genres in top 10 movies for each latent factor .

It is also worth noting that although the number of genres and the number of latent factors are both 20. There is no exact one-to-one relationship between genres and latent factors. This is because NNMF calculates implied styles of the movies, and there could be more factors influencing it other than the movies' genre.

Question 24-25

Design a MF with bias collaborative lter to predict the ratings of the movies in the MovieLens dataset and evaluate it's performance using 10-fold cross-validation. Sweep k (number of latent factors) from 2 to 50 in step sizes of 2, and for each k compute the average RMSE and average MAE obtained by averaging the RMSE and MAE across all 10 folds. 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.
Use the plot from question 24, to nd 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.

```
Best k for MAE is:  4 MAE =  0.7059363877940632
Best k for RMSE is:  4 RMSE =  0.9252020075385389
```

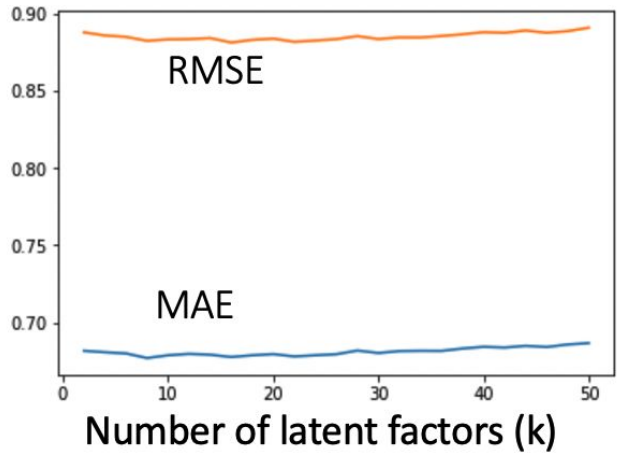


Question 26-28

Design a MF with bias collaborative lter to predict the ratings of the movies in the popular/unpopular/high-variance movie trimmed test set and evaluate it's performance using 10-fold cross validation. Sweep k (number of latent factors) from 2 to 50 in step sizes of 2, and for each k compute the average RMSE obtained by averaging the RMSE across all 10 folds. Plot average RMSE (Y-axis) against k (X-axis). Also, report the minimum average RMSE

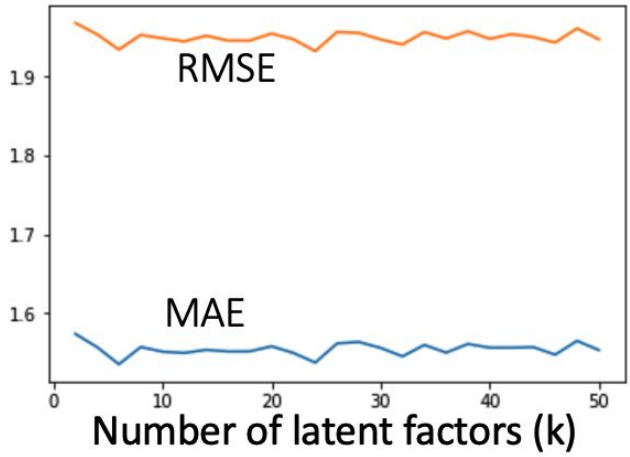
Popular Movies:

```
Popular Movies
Best k for MAE is:  8 MAE =  0.6770409043742929
Best k for RMSE is:  16 RMSE =  0.8812395875576818
```



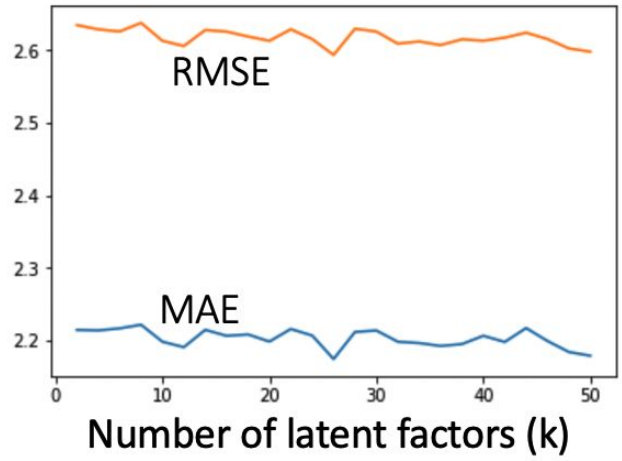
Unpopular Movies

```
Unpopular Movies
Best k for MAE is:  6 MAE =  1.5352741885698558
Best k for RMSE is:  24 RMSE =  1.9322396842250213
```



High-variance Movies

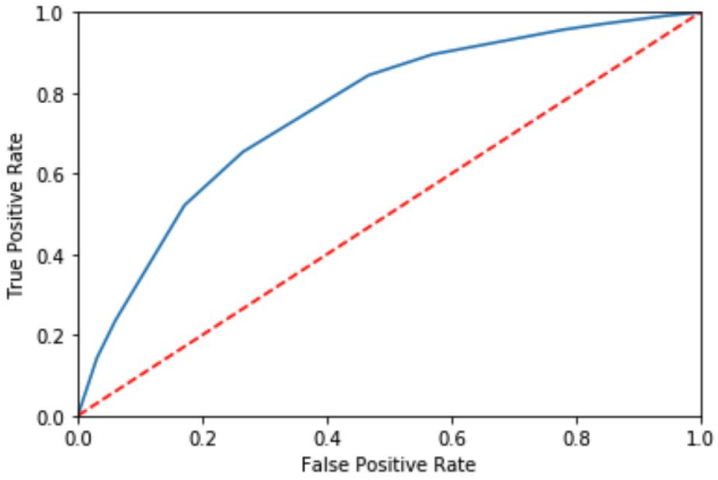
```
High-variance Movies
Best k for MAE is:  26 MAE =  2.1741789359603083
Best k for RMSE is:  26 RMSE =  2.5934378349570033
```



Question 29

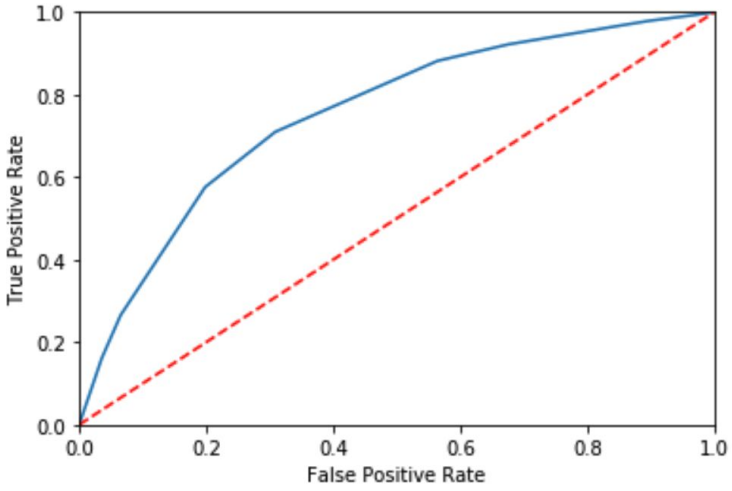
Plot the ROC curves for the MF with bias collaborative filter designed in question 24 for threshold values [2:5; 3; 3:5; 4]. For the ROC plot-ting use the optimal number of latent factors found in question 25. For each of the plots, also report the area under the curve (AUC) value.

Threshold = 2.5



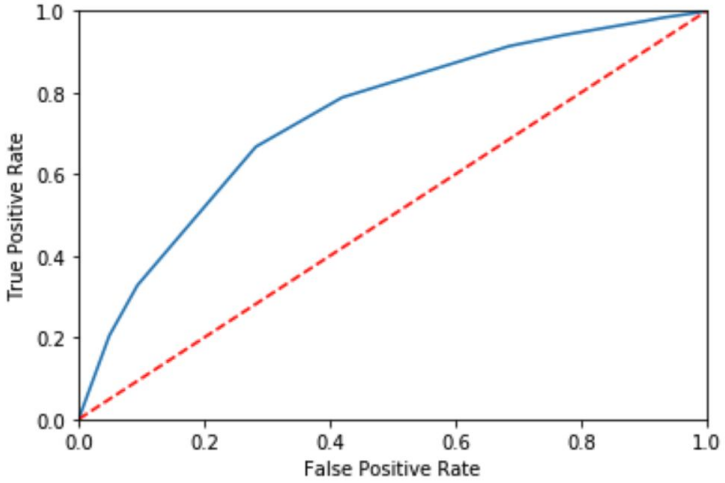

```
AUC =  0.7559713318289734
```

Threshold = 3.0



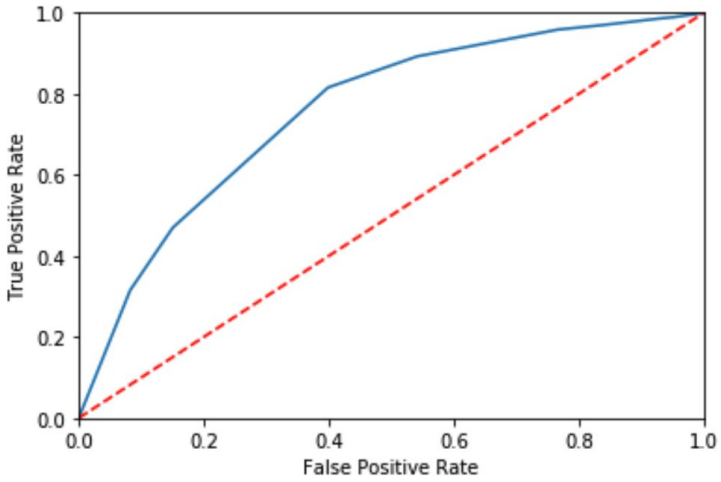

```
AUC =  0.7521800858429373
```

Threshold = 3.5




```
AUC =  0.7385882352941177
```

Threshold = 4.0




```
AUC =  0.7578266705452151
```

Question 30

Design a naive collaborative filter to predict the ratings of the movies in the MovieLens dataset 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: 0.9409907273946734

Question 31

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: 0.9374659905180754

Question 32

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: 0.8967892011507379

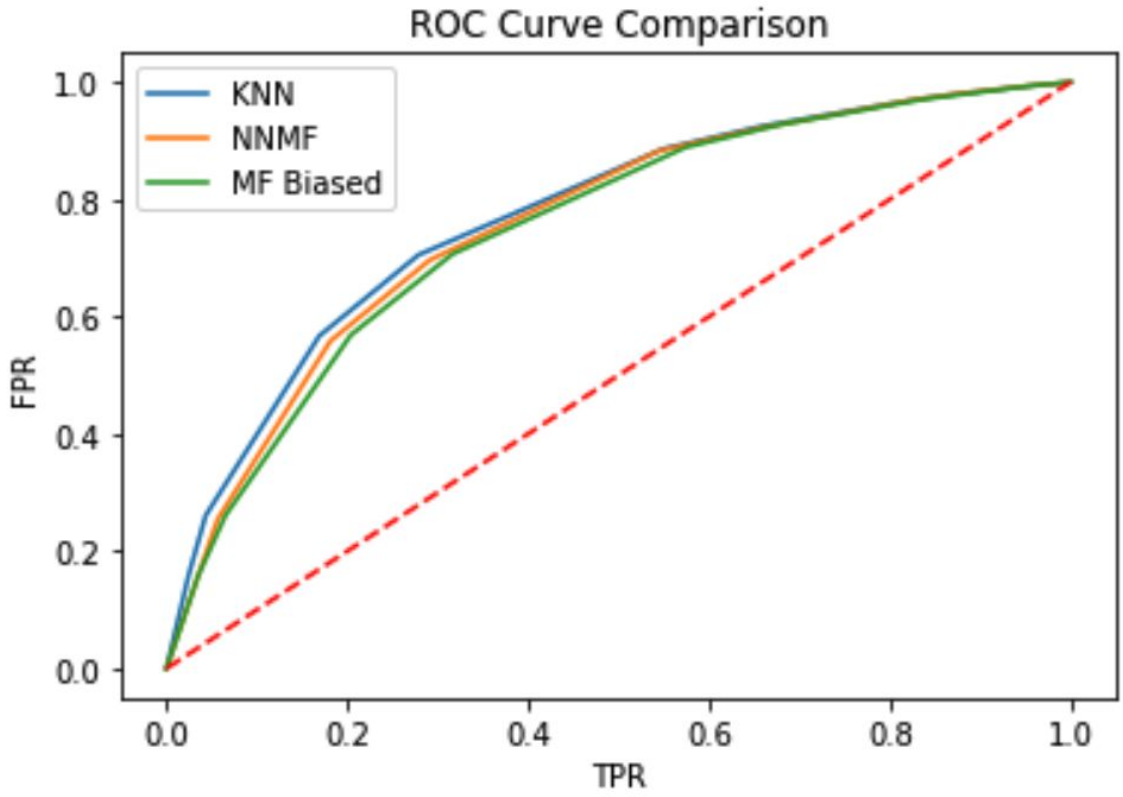
Question 33

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: 1.3969862694976212

Question 34

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.



The performances of the three algorithms are very close. At TPR level of 0.6 or higher, the three algorithms almost perform the same. At TPR level of less than 0.6, k-NN is the best, NNMF is

the second and MF with bias is the last. k-NN outperforms NNMF by about 0.03 and NNMF outperforms MF with bias by about 0.03 as well.

Question 35

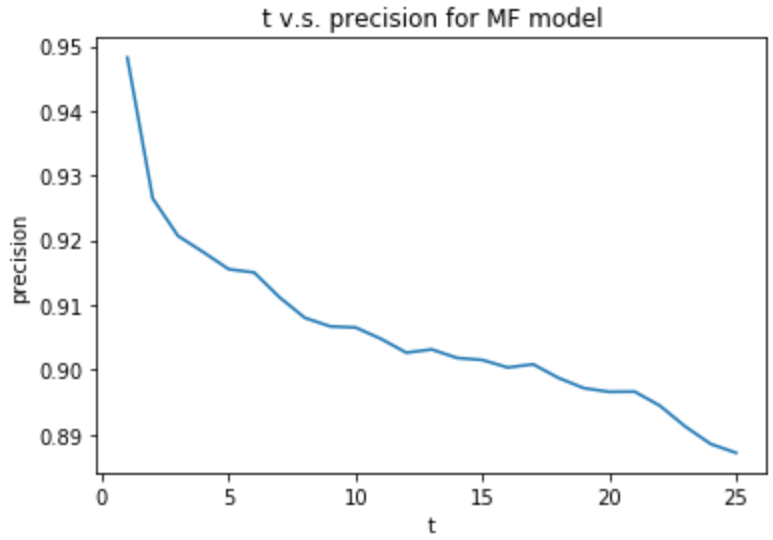
Precision and Recall are defined by the mathematical expressions given by equations 12 and 13 respectively. Please explain the meaning of precision and recall in your own words.

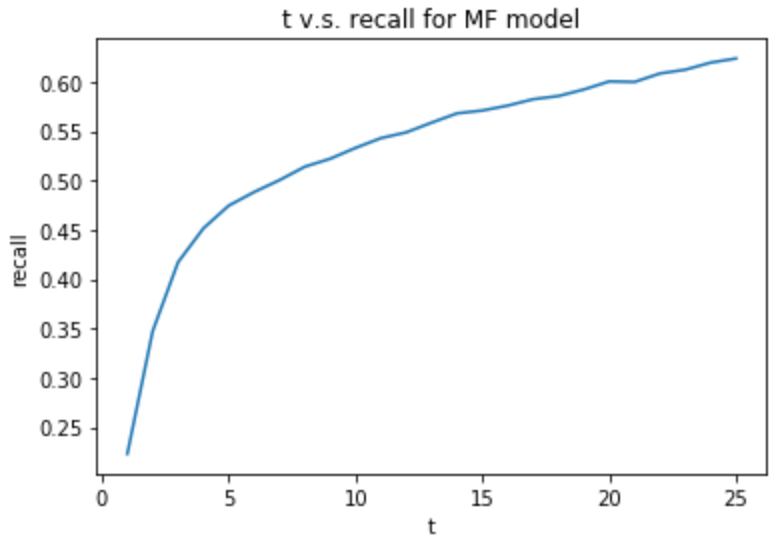
Precision: The percentage of recommended items that are liked by the users. This measures how many recommned items are liked by the users.

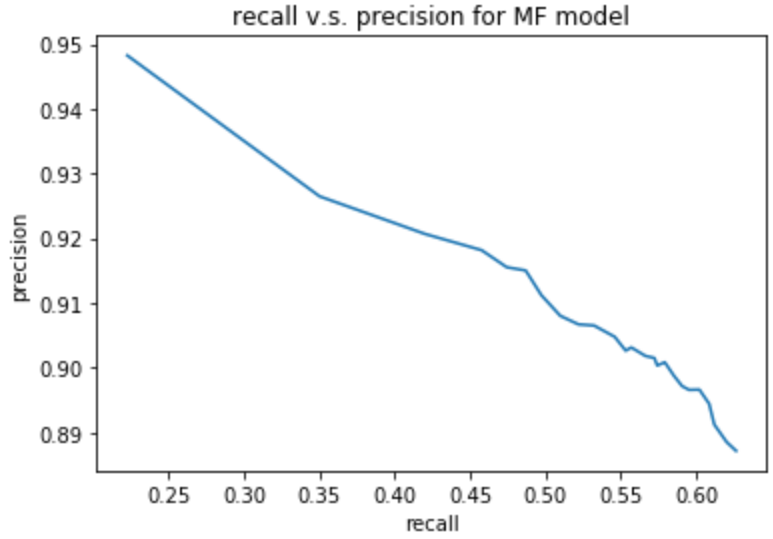
Recall: The percentage of the liked items that are recommended to the users. This basically measures how many liked items have been recommended.

Question 36

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.







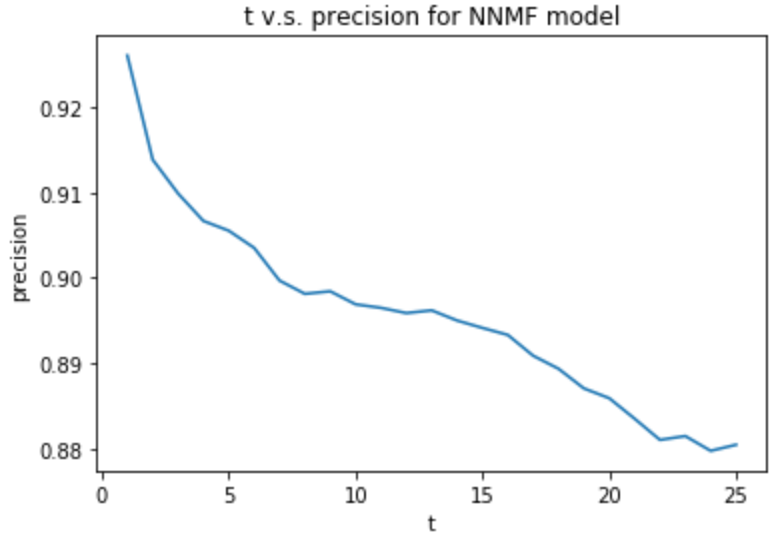
Precision v.s. t plot: The plot is not monotonic, but as you include more times into the plot, it is less likely to maintain a very high precision.

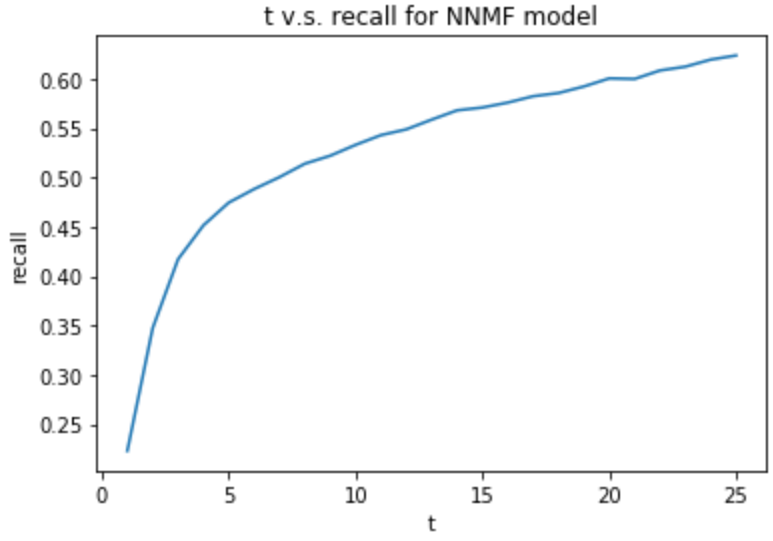
Recall v.s. t plot: In contrast, if you include more items, it is more likely you will include some items that are liked by the users. If you include all the items the recall will be 1. Therefore, the recall curve is monotonically increasing.

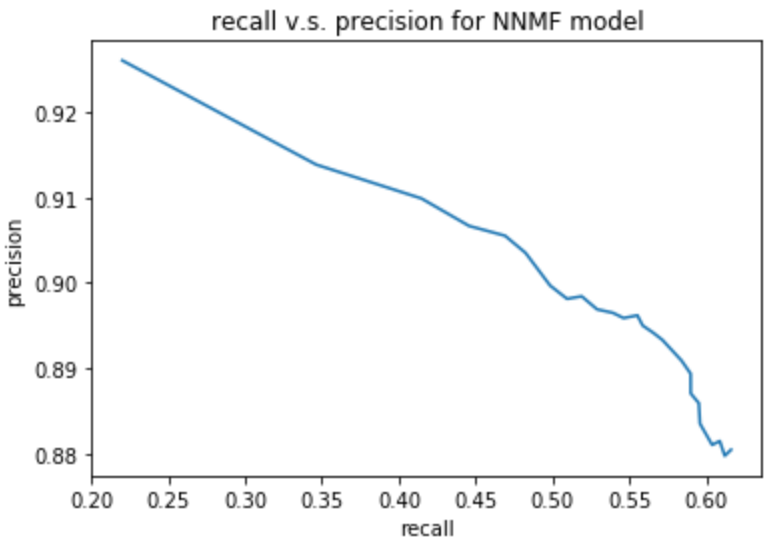
Precision v.s. recall plot: There is an inverse linear relationship between precision and recall. The value of precision decreases, when precision increases.

Question 37

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.







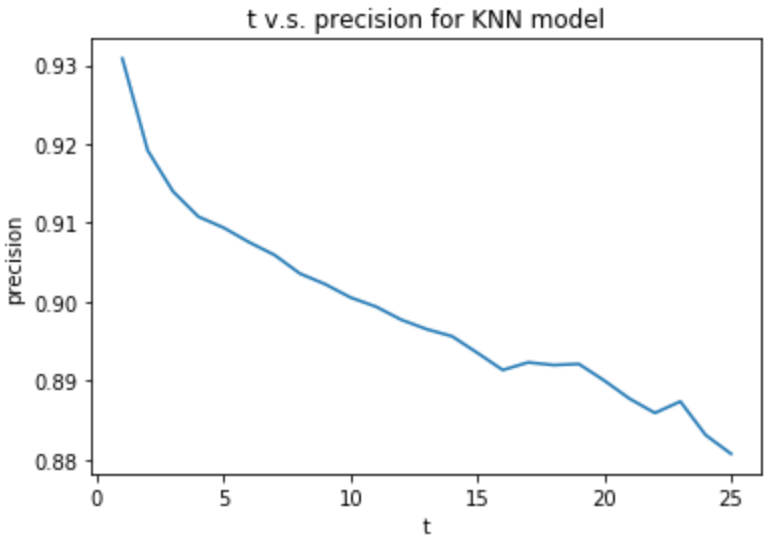
Precision v.s. t plot: The plot is not monotonic, but as you include more items into the plot, it is less likely to maintain a very high precision.

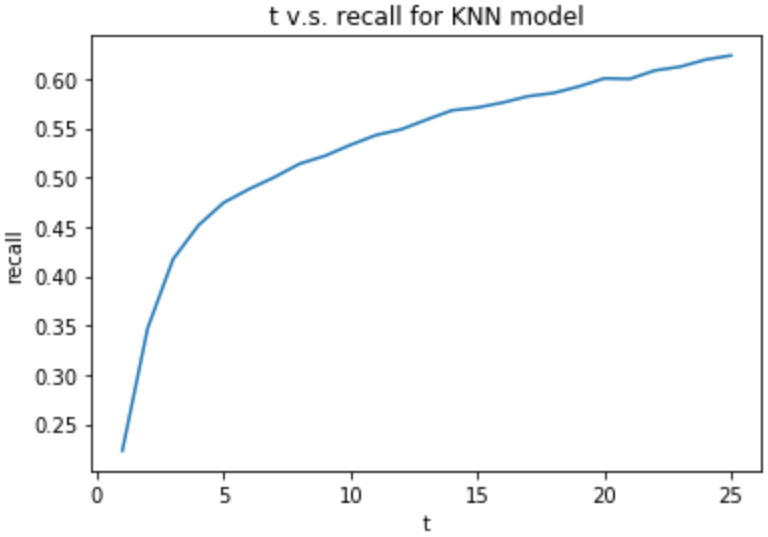
Recall v.s. t plot: In contrast, if you include more items, it is more likely you will include some items that are liked by the users. If you include all the items the recall will be 1. Therefore, the recall curve is monotonically increasing.

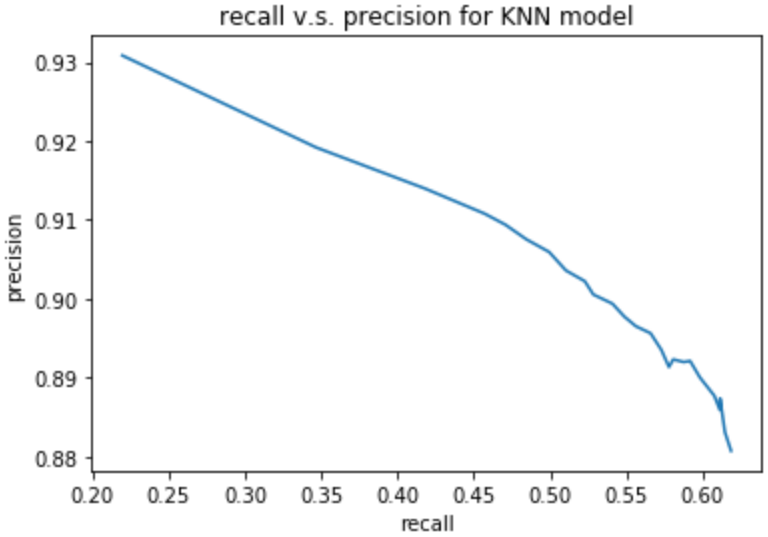
Precision v.s. recall plot: There is an inverse relationship between precision and recall. The value of precision decreases, when precision increases

Question 38

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.


t v.s. precision for KNN model
precision
t
0.93
0.92
0.91
0.90
0.89
0.88
0
5
10
15
20
25


t v.s. recall for KNN model
recall
t
0.60
0.55
0.50
0.45
0.40
0.35
0.30
0.25
0
5
10
15
20
25


recall v.s. precision for KNN model
precision
recall
0.93
0.92
0.91
0.90
0.89
0.88
0.20
0.25
0.30
0.35
0.40
0.45
0.50
0.55
0.60

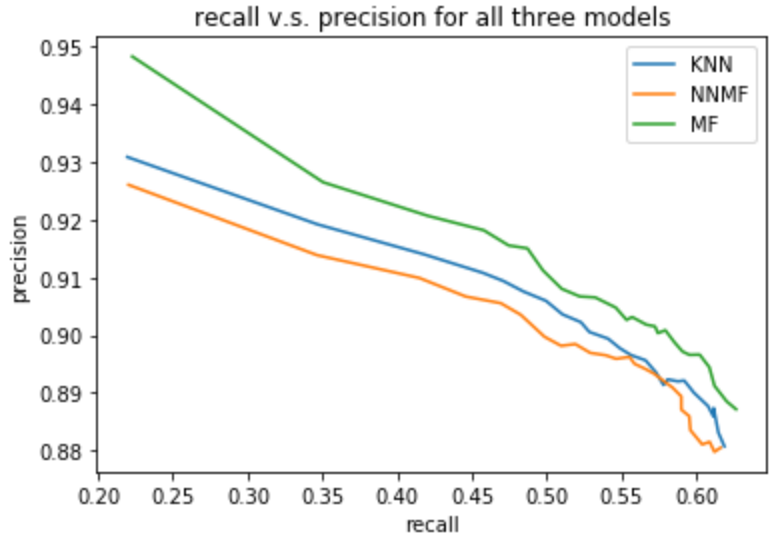
Precision v.s. t plot: The plot is not monotonic, but as you include more items into the plot, it is less likely to maintain a very high precision.

Recall v.s. t plot: In contrast, if you include more items, it is more likely you will include some items that are liked by the users. If you include all the items the recall will be 1. Therefore, the recall curve is monotonically increasing.

Precision v.s. recall plot: There is an inverse relationship between precision and recall. The value of precision decreases, when precision increases.

Question 39

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.



Both three curves share the similar trends, and the MF curve tend to have better precision and recall curve. Therefore, the MF is more suitable for the ranking task of the given data.