Project 3 Report

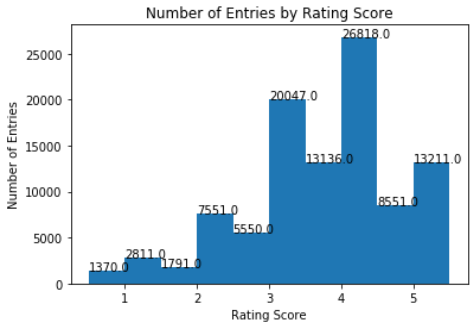
Mahmoud Essalat, Sherry He, Weitang Sun, Siqi Huang
ID: 005034839, 805040110, 904946260, 504490530

Question 1.

The sparsity is 0.017. It means that only a small fraction of the ratings is available and thus the rating data is pretty sparse. The sparsity is a problem we need to deal with for the recommendation system.

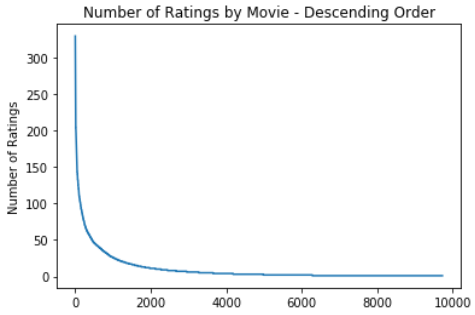
Question 2.

The histogram showing the frequency of the rating values is displayed in the following. We can see that the rating distribution is skewed left. Most people will give a rating ranging 3 to 5. They might have a bias towards high rating values. Also, it is not a monotonically decreasing distribution.



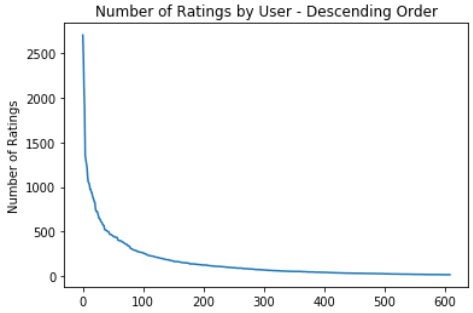
Question 3.

The plot is showing in the following figure. We can see that most movies (more than 90%) receives less than 50 ratings while very few movies review more than 100 ratings. But the important fact is that the distribution has a long tail/ heavy tail. So it represents nonlinear factors in the data that has been gathered. As a result, this cannot be estimated with linear methods like recommendation systems (NMF, etc).



Question 4.

The plot is showing in the following. We can see that most users give less than 500 ratings while very few users give more than 1000 ratings. The distribution has a long tail like question 3.

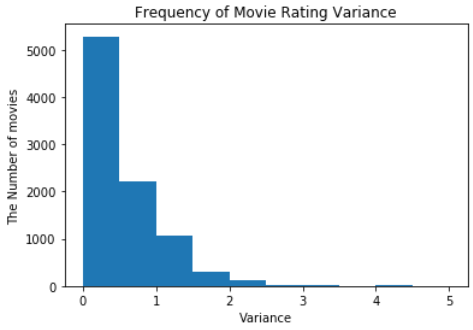


Question 5.

The fact that the above distributions display the long-tail feature has important implications for neighborhood-based collaborative filtering algorithms. Neighborhood-based collaborative are usually building on the assumption that movies are rated frequently. However, in this case, because the available ratings are sparse, some movies (highly rated ones) may not be able to represent others (lowly rated ones). There might be underlying factors in differentiating these two kinds of movies. Therefore, without handling this missing information, the prediction results might be biased. The recommendation algorithm needs to accommodate the sparsity problem and handle the long-tail distribution. . The heavy tail feature, it represents nonlinear factors in the data that has been gathered. As a result, this cannot be estimated with linear methods like recommendation systems (NMF, etc).

Question 6.

The plot is showing in the following. It shows that most movies have variances less than 1, and very few have variance more than 2.



Question 7.

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

It simply means adding the ratings that are available divided by number of ratings.

Question 8.

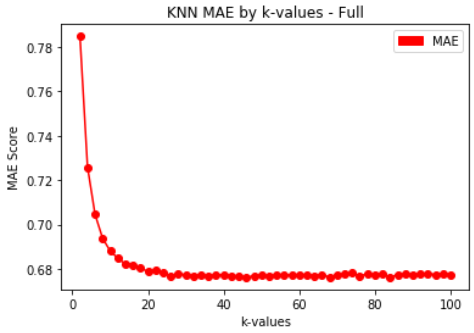
$I_u \cap I_v$ denotes the set of movies that have been rated by both users u and v. In our case, as the ratings is sparse, it is likely $I_u \cap I_v = \emptyset$.

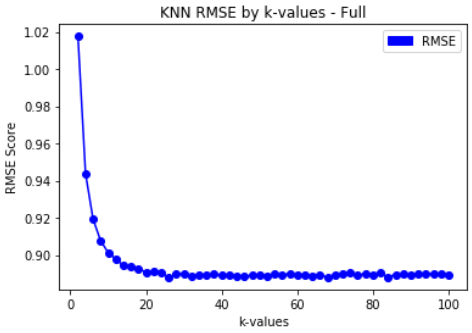
Question 9.

It is because different users have different rating scales. Some may consistently rate highly and some may consistently rate poorly. By using the mean-centered ratings, we eliminated this individual bias to avoid misleading predictions.

Question 10 & 11.

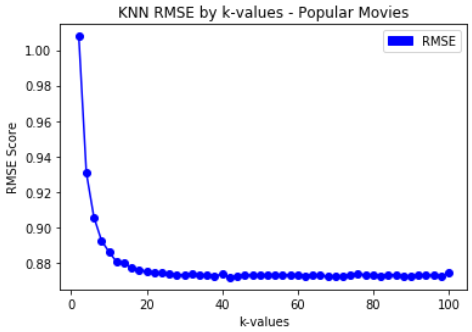
The plots are showing in the following. The lines become steady around k = 25 so the minimum k is 25. The steady state values are 0.67 for MAE and 0.89 for RMSE respectively.





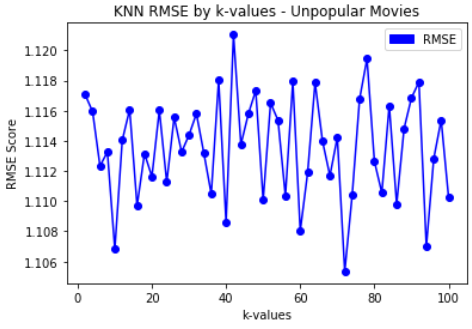
Question 12.

The minimum average RMSE is 0.87.



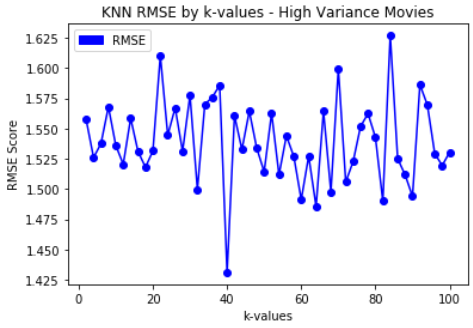
Question 13.

The minimum average RMSE is 1.10.



Question 14.

The minimum average RMSE is 1.43.

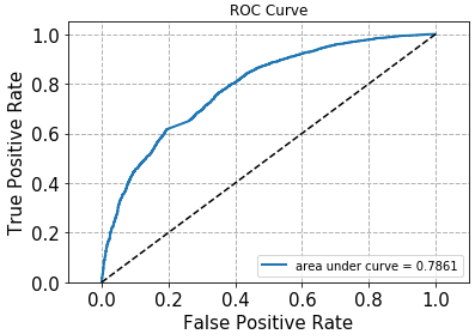


Summary of Question 12-14: *k*-NN collaborative filter performs the best in predicting the rating of the movies in the popular movie trimming case and performs the worst in predicting the ratings of the movies in the high variance movie trimming case. Also it performs very poorly in the unpopular movies case. The reason is that, for the popular movie case, there is enough information to do the prediction. However, for the high variance movie case, the movies receive very different ratings already and thus it is hard to predict the rating. In other words the unpopular movies or high variance movies have a heavy tail distribution with chance of tail ratings to be relatively higher than other distributions. So it shows there are some nonlinear factors in the unpopular or high variance movies which cannot be estimated with linear estimations like KNN, NMF , SVD, etc.

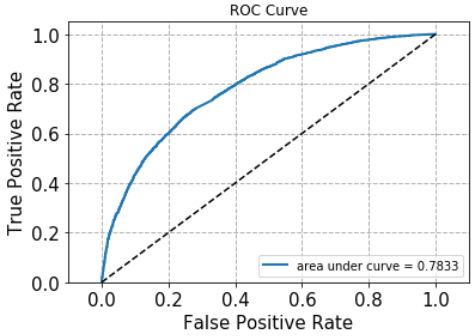
Question 15.

The plots are showing in the following. We use $k = 25$ as illustrated in Question 10&11. As the area (under the ROC curve) represents the measure of the quality of the recommendation system, we can see that when threshold is 3, the corresponding area is the largest and thus k-NN collaborative filter performs the best. Overall, the k-NN collaborative filers perform moderately.

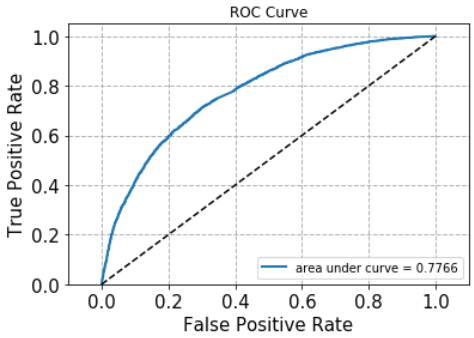
Threhold = 2.5



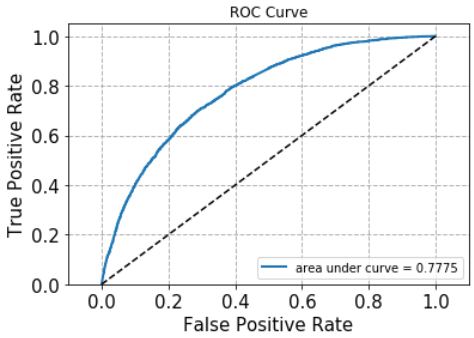
Threhold = 3.0



Threhold = 3.5



Threhold = 4.0

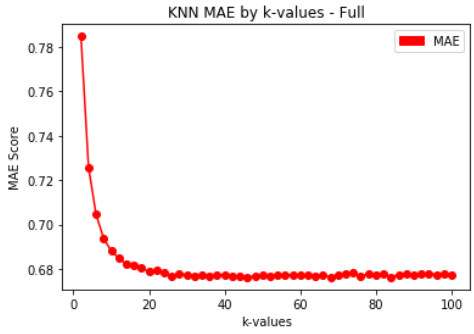


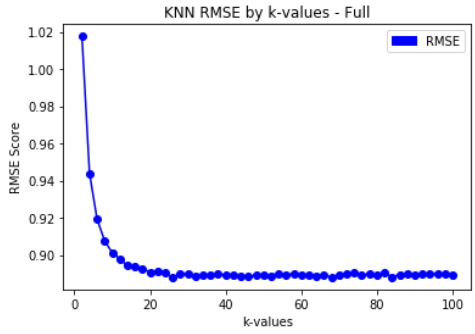
Question 16.

The equation 5 is $\underset{U,V}{\text{minimize}} \sum_{i=1}^{m} \sum_{j=1}^{n} W_{ij} \, (r_{ij} - (UV^T)_{ij})^2$. It is non-convex in U, V because of the term $(UV^T)_{ij}$. For fixed U, the equation could be formulated as a least-squares problem:

$$\underset{V}{\text{minimize}} \sum_{(i,j)|r_{ij}\, known} (r_{ij} - (UV^T)_{ij})^2$$

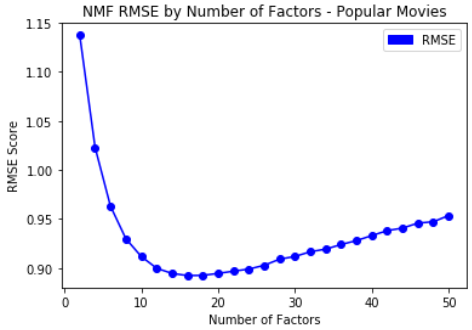
Question 17.
The plots are showing in the following.



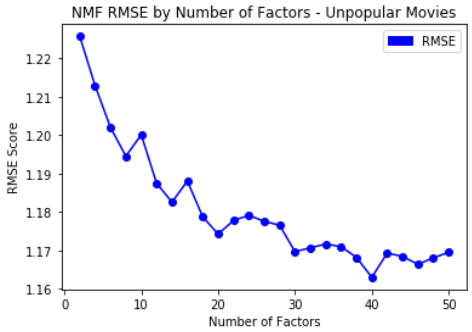


Question 18.
From plots in the previous question and by printing out the optimal number of latent factors, the optimal number of latent factors from RMSE plot is 16 and from MAE plot is 22. The minimum average RMSE is 0.913 and the minimum average MAE is 0.693. There are 19 movie genres so the optimal number of latent factors is basically the same as the number of movie genres. The 19 movie genres are listed in the following: Action, Adventure, Animation, Children, Comedy, Crime, Documentary, Drama, Fantasy, Film-Noir, Horror, IMAX, Musical, Mystery, Romance, Sci-Fi, Thriller, War, Western.

Question 19.
The minimum average RMSE is 0.89.

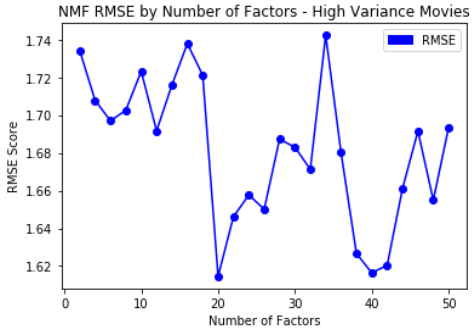


Question 20.
The minimum average RMSE is 1.16.



Question 21.

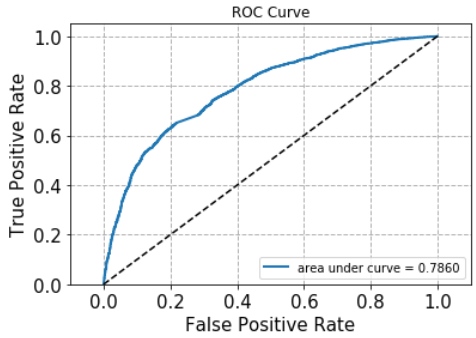
The minimum average RMSE is 1.61.



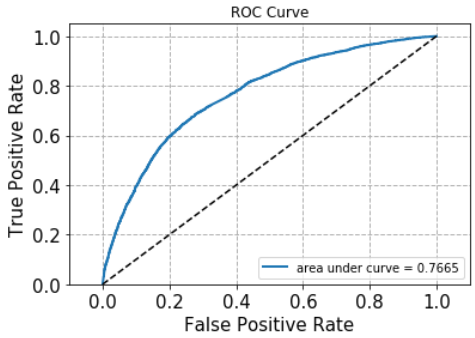
Summary of Question 19-21: NNMF collaborative filter predicts the best in the popular movie case and predicts the worst in the high variance movie case. The reason is that, for the popular movie case, there is more information to do the prediction. However, for the high variance movie case, the movies receive very different ratings already and thus it is hard to predict the rating. . Also it performs poorly in the unpopular movies case. The reason is that, for the popular movie case, there is enough information to do the prediction. However, for the high variance movie case, the movies receive very different ratings already and thus it is hard to predict the rating. In other words the unpopular movies or high variance movies have a heavy tail distribution with chance of tail ratings to be relatively higher than other distributions. So it shows there are some nonlinear factors in the unpopular or high variance movies which cannot be estimated with linear estimations like KNN, NMF , SVD, etc.

Question 22.

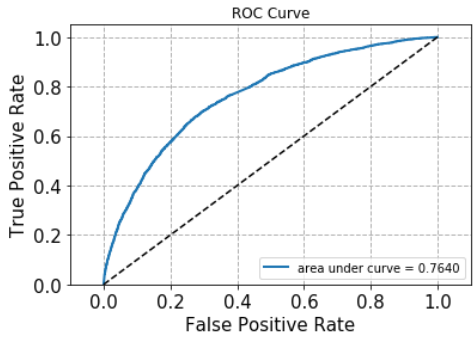
Threhold = 2.5



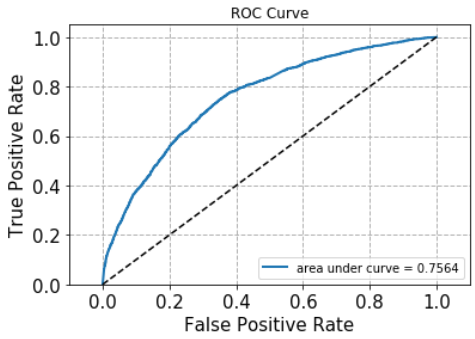
Threhold = 3.0



Threhold = 3.5



Threhold = 4.0



Compare with the previous plots, we can see that *k*-NN and NNMF have similar performance.

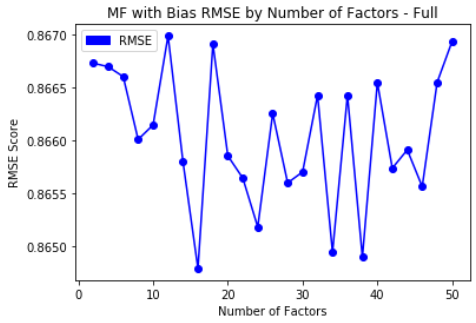
Question 23.

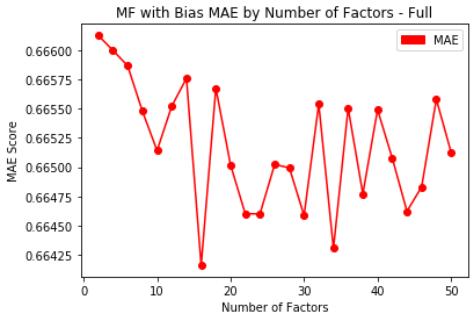
We listed the 20 columns of V in the following table. We can see that each column groups similar movie genres together. For example, group $k = 0$ captures crime and drama most, group $k = 1$ highlights romantic comedy movies, and group $k = 9$ is strongly tied to movies in the genre of horror and action.

====== k : 0 Drama Drama|Romance Drama|Thriller Comedy|Drama|Romance Crime|Film-Noir Crime|Drama|Mystery|Romance|Thriller Comedy Comedy|Drama|Romance Crime|Thriller Crime|Drama|Mystery|Romance|Thriller	====== k : 1 Action|Crime|Drama|Sci-Fi|Thriller Crime|Drama Drama|Romance Comedy Comedy|Crime Comedy|Romance Comedy|Drama|Thriller Comedy|Drama Drama|Romance|War Action|Comedy|Romance	====== k : 2 Comedy Adventure|Western Drama|Fantasy Drama|Sci-Fi Adventure|Fantasy|Thriller|IMAX Crime|Drama Adventure|Fantasy|Romance|Sci-Fi|Thriller Action|Crime|Drama Drama Action|Adventure|Drama|Fantasy|Thriller	====== k : 3 Comedy|Drama Crime|Film-Noir|Mystery Drama Action|Comedy|Crime|Drama Comedy|Romance Comedy Drama|Horror|Mystery|Thriller Drama Drama Action|Crime|Drama|Horror	====== k : 4 Action|Comedy|Crime|Thriller Documentary Comedy|Fantasy Drama|Romance Drama|Thriller Comedy Drama|Romance Crime|Drama Comedy|Horror|Mystery|Thriller Action|Adventure|Fantasy|Sci-Fi
====== k : 5 Drama Comedy|Drama|Mystery Fantasy|Western Action|Crime|Drama|Thriller Drama|Romance Action|Fantasy|Sci-Fi|Thriller Comedy Comedy|Drama Fantasy Drama|Thriller	====== k : 6 Action|Crime|Thriller|Western Drama|Thriller Horror|Mystery Comedy|Horror|Thriller Action Drama|Romance Action|Crime|Drama|Sci-Fi|Thriller Action|Comedy|Drama|Horror|Thriller Drama|Romance Children|Musical	====== k : 7 Action|Comedy|Crime|Romance Comedy|Romance Comedy|Drama|Thriller Drama Horror Drama|Mystery Action|Adventure|Sci-Fi Comedy|Drama|War Crime|Drama Action|Comedy|Crime|Thriller	====== k : 8 Comedy|Drama|Romance Drama|War Drama|Romance Comedy|Fantasy|Romance Action|Adventure|Comedy|Fantasy|Mystery Action|Drama|Romance Comedy|Drama Crime|Drama|Romance Sci-Fi Adventure|Sci-Fi	====== k : 9 Action|Comedy Crime|Horror|Thriller Drama Comedy Horror Horror|Sci-Fi Action|Comedy|Crime|Romance Drama Action|Adventure|Drama|Sci-Fi|Thriller Action|Fantasy|Sci-Fi|Thriller
====== k : 10 Drama|War Comedy|Fantasy Comedy|Drama|Romance Drama|Thriller Crime|Drama|Mystery|Thriller Action|Crime|Sci-Fi Action|Comedy Action|Drama|Mystery|Sci-Fi|Thriller|IMAX Musical Drama|Romance	====== k : 11 Horror Comedy Comedy|Drama Drama|Thriller Drama Action|Crime|Drama|Thriller Crime|Drama|Thriller Comedy Action|Comedy Comedy	====== k : 12 Action|Adventure|Comedy|Crime Action|Comedy Drama|Fantasy|Mystery|Romance Action|Adventure|Sci-Fi Adventure|Comedy Drama|Romance Comedy|Musical|Romance Comedy Comedy|Romance Drama|Horror|Sci-Fi|Thriller	====== k : 13 Comedy|Crime|Thriller Action|Comedy Comedy|Drama Drama|Fantasy|Musical Drama|Thriller Drama Comedy|Horror|IMAX Crime|Drama Action|Adventure|Sci-Fi|Thriller Horror	====== k : 14 Action|Children|Comedy Comedy|Drama Drama|Romance|War Horror Action|Horror|Sci-Fi Comedy|War Horror Action|Comedy|Fantasy|Horror|Thriller Comedy Comedy|Drama|Romance
====== k : 15 Crime|Drama|Thriller Action|Adventure|Drama|Sci-Fi|Thriller Adventure|Animation|Children|Comedy Animation|Children|Fantasy|Mystery Comedy Action|Fantasy|Sci-Fi|Thriller Comedy|Sci-Fi Comedy|Drama|Romance Action|Adventure|Crime|Thriller Crime|Drama	====== k : 16 Crime|Horror|Mystery|Thriller Comedy|Romance Comedy|Musical Documentary Drama|Fantasy|Horror Action|Comedy Comedy|Musical|Romance Drama Sci-Fi Comedy	====== k : 17 Adventure|Drama|Thriller Action|Crime|Thriller Comedy Drama Drama Drama|Romance|War Drama|Musical Action|Crime|Thriller Comedy|Drama|Romance Adventure	====== k : 18 Drama Action|Comedy|Crime|Thriller Adventure|Animation|Children|Fantasy Comedy Comedy|Drama Comedy|Drama Comedy|Crime|Drama Comedy Drama|Romance Comedy|Drama	====== k : 19 Action|Comedy|Horror|Sci-Fi Drama Comedy|Romance Comedy|Drama|Musical Drama|Thriller Adventure|Children|Drama|Fantasy|IMAX Adventure|Animation|Comedy Drama Comedy|Romance Comedy|Horror|IMAX

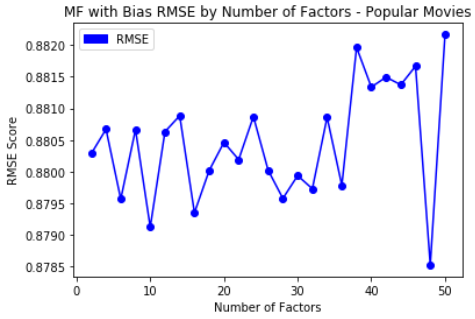
Question 24 & 25.

The plots are showing in the following. The optimal number of latent factors is 16. The corresponding minimum average RMSE and MAE are 0.86 and 0.66 respectively.

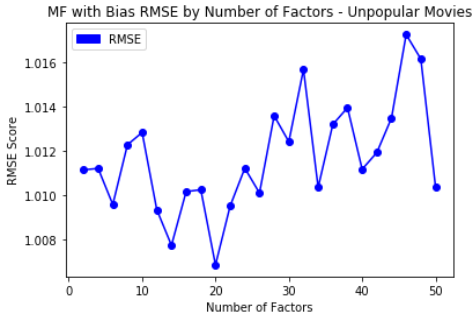




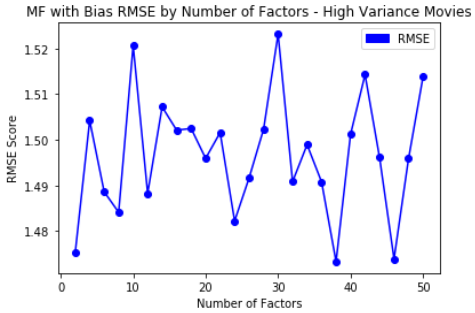
Question 26.
The minimum average RMSE is 0.88.



Question 27.
The minimum average RMSE is 1.01.



Question 28.
The minimum average RMSE is 1.47.

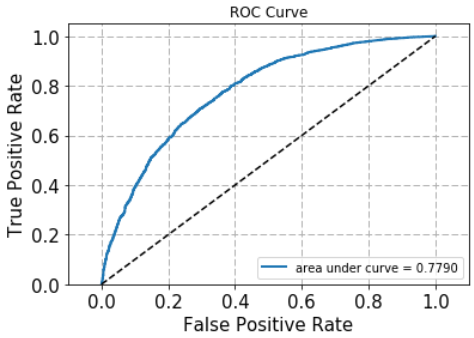


Summary of Question 26-28: MF with bias collaborative filter predicts the best in the popular movie case and predicts the worst in the high variance movie case. MF with bias collaborative filter performs similar to *k*-NN and better than NNMF collaborative filter, in terms of average RMSE.

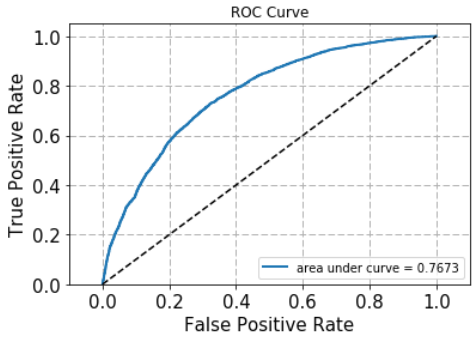
Question 29.

The number of latent factors we use is 16. MF with bias based collaborative filter has a modest performance, compared with *k*-NN and NNMF.

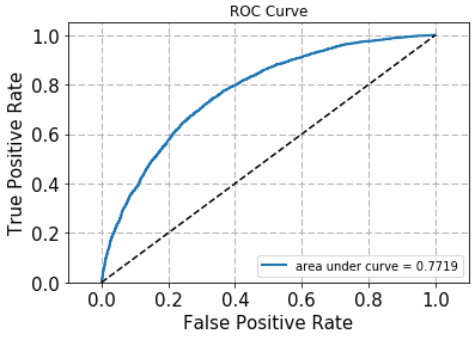
Threhold = 2.5



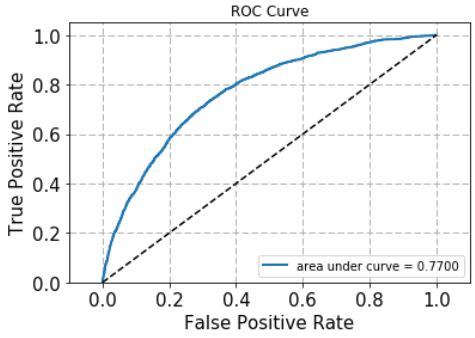
Threhold = 3.0



Threhold = 3.5



Threhold = 4.0

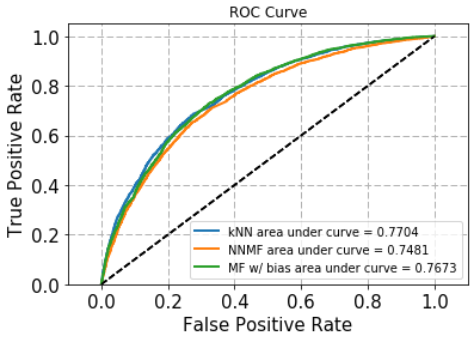


Question 30-33.

Native Collaborative filter	Average RMSE
All Movies	0.93
Popular Movie Trimmed test set	0.96
Unpopular Movie Trimmed test set	0.97
High Variance Movie Trimmed test set	1.01

From the above table, we can see that High Variance setting still has the highest average RMSE while all movies setting has the lowest. The pattern is similar to the previous ones. However, with Native Collaborative filter, the average RMSE performances from different settings do not share the same large differences as other filters applied previously. With Native Collaborative filter, average RMSE from all settings performs similar to the unpopular movie setting from the k-NN filter, NNMF collaborative filter and MF with bias collaborative filter. It is because for unpopular movies, there is less information to make predictions and thus using the mean (naïve filter) predicts better.

Question 34.

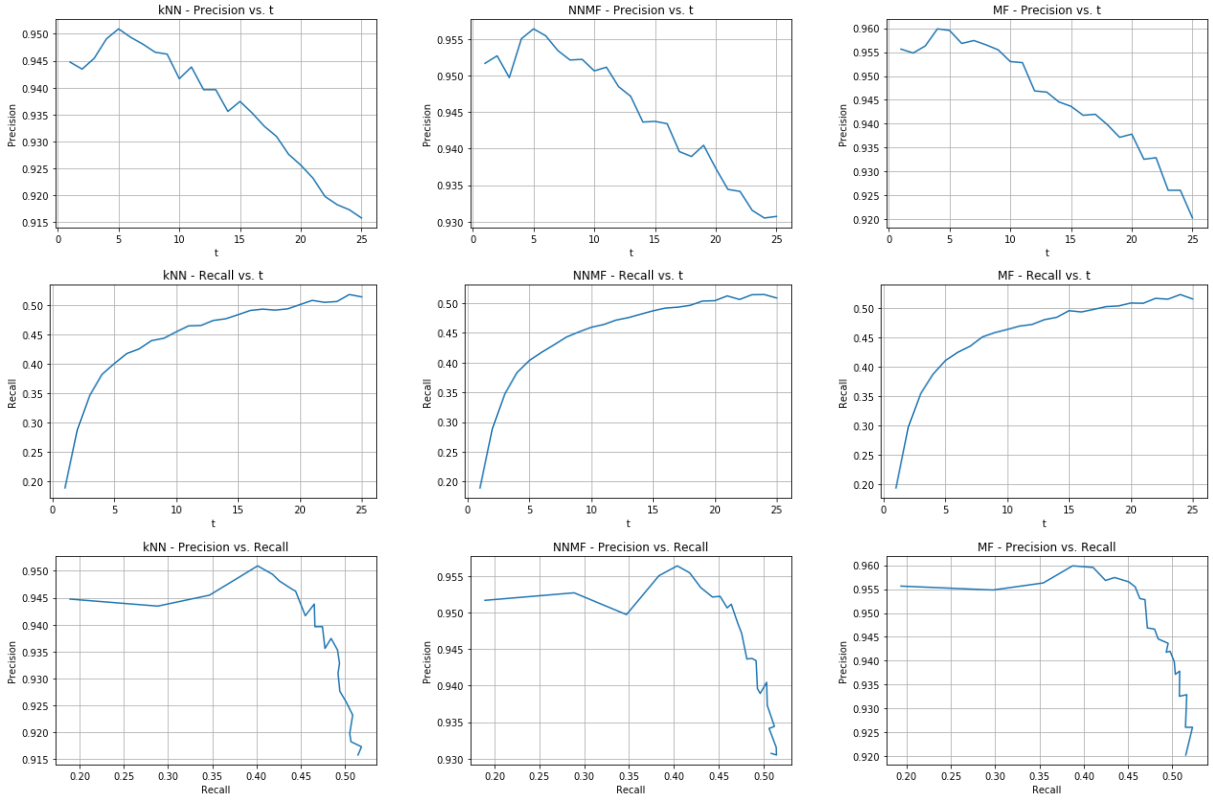


From the above ROC curve plot, we can see that all three filters perform similarly, as three lines are close to each other. However, NNMF performs the worst. The NNMF ROC curve is dominated by the other two ROC curves. Therefore, both *k*-NN and MF with bias filters are best suited for movie rating prediction in this case.

Question 35.

Precision is the percentage of suggested items of size *t* recommended to the user that is liked by the user. Recall is the percentage of items liked by the user that is in the set of items of size *t* recommended to the user.

Question 36, 37 & 38.



We show the Prevision v.s. Size t, Recall v.s. Size t and Precision v.s. Recall plots from k-NN, NNMF and MF algorithms in the above.

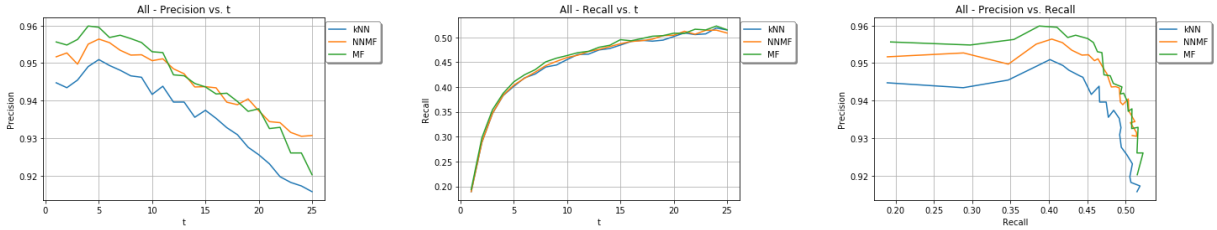
The precision plots show that the average precision is not monotonic in t. Overall, it is decreasing with t but there are bumps here and there. The reason is that, from the definition of precision, $|S(t) \cap G|$ in the numerator and $S(t)$ in the denominator are both functions of t and may change with t in a different way.

The recall plots show that the average recall is almost monotonically increasing in t. The reason is that, from the definition of recall, $|S(t) \cap G|$ in the numerator is a function of t, $|G|$ is a fixed number and the numerator increases with t.

The Precision v.s. Recall plots show the tradeoff between average precision and average recall. The increase of recall does not change precision much initially with a huge drop of precision later.

The shapes of the plots for k-NN, NNMF-based and MF with bias-based collaborative filters are similar to each other.

Question 39.



The above plots show a direct performance comparison among three filters. MF almost dominates NNMF and *k*-NN entirely. *k*-NN filter is user-based, so it is easy to implement. NNMF is latent-factor based and shares similar performance with *k*-NN. MF with bias based collaborative filters learned the user-specific bias and item-specific bias. This additional information helps with the predictions and therefore MF with bias based collaborative filter works the best in the movie rating and ranking case.