ECE-219

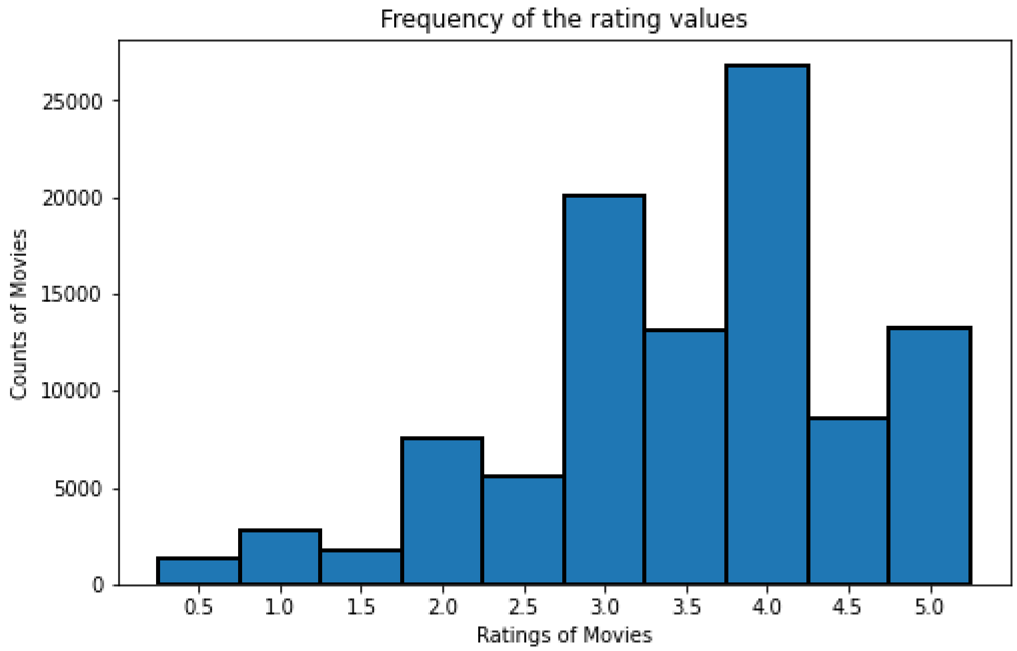
Data Representation and Clustering

Team Member Names:

Tianpei Gu, 405863048
Yuxin Huang, 105711853
Yilin Xie, 405729012

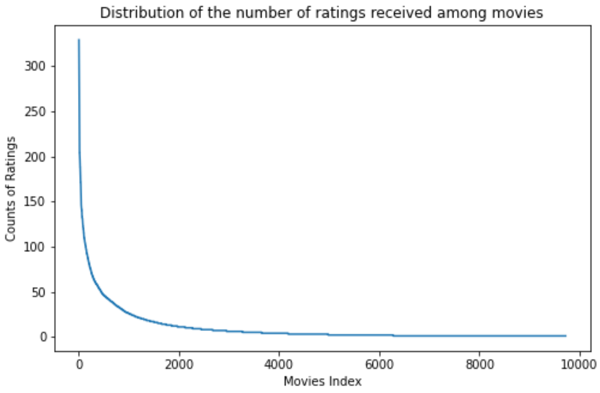
Question 1

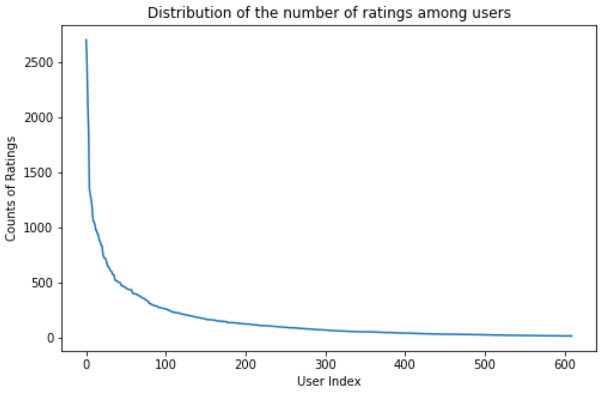
- `Sparsity: 0.016999683055613623`



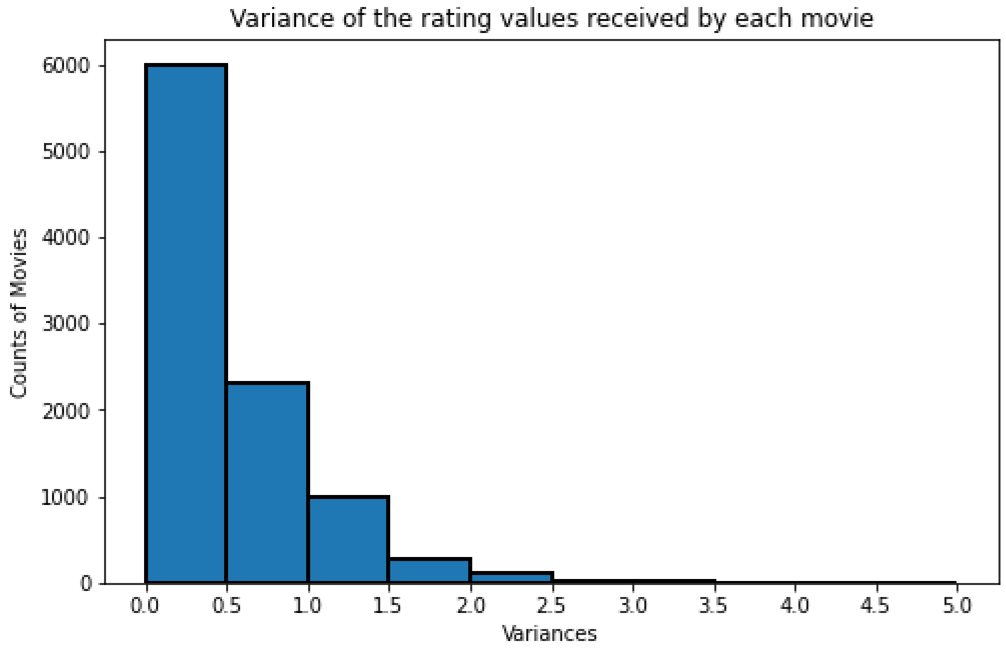
\

- Most movies have comparatively high ratings, and only a small percentage of movies have bad ratings.





- Based on the distribution of the number of ratings among **movies and users**, we can see that most movies do not receive a lot of ratings but only a few movies received the majority of the ratings; most users do not rate a lot of movies but only a few of them rated a lot of movies. It proves the sparsity of the ratings matrix. Therefore, for the recommendation process, further cleaning or regularization are required to prevent overfitting.



- Most movies have small rating variance ranging from 0 to 2, which means that most of the users have the same opinion towards each movie.

Question 2

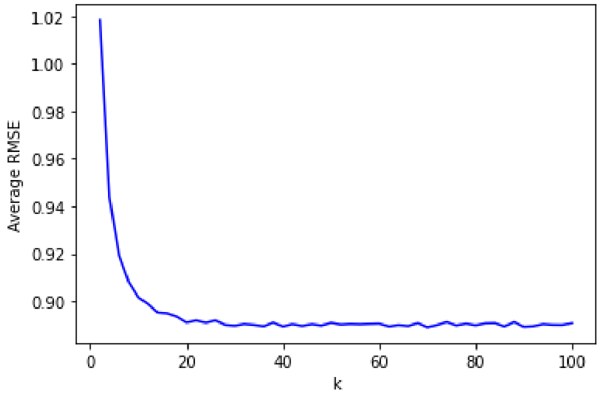
- $\mu_u = \frac{\Sigma_{k \in I_u} r_{uk}}{\left| I_u \right|}$
- $I_u \cap I_v$ means Set of item indices for which ratings have been specified by both user u and v. It can be an empty set, because a movie can be neither rated by u nor v.

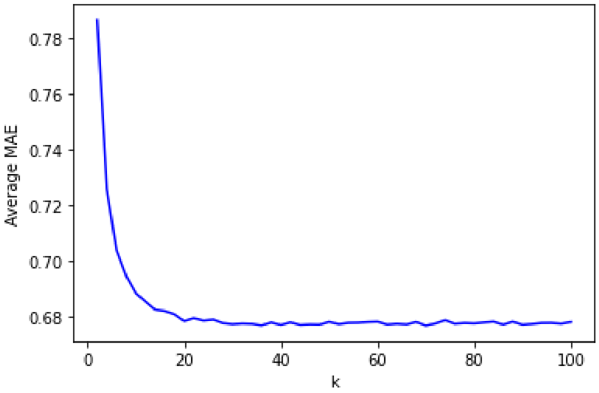
Question 3

- Mean-centering will alleviate the impact of extreme users and reduce bias (e.g. users who either rate all items highly or rate all items poorly). It helps to train a more accurate model.

Question 4

- Our results are as follows:
  - Both average RMSE and MAE monotonically decrease as k increases.


k-NN collaborative filter with 10-fold cross validation.


k-NN collaborative filter with 10-fold cross validation.

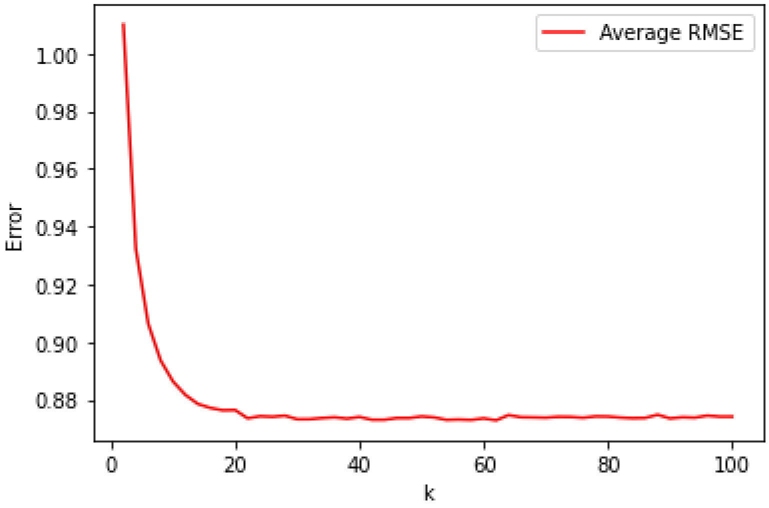
Question 5

- Minimum k for which RMSE converges: **16**, with average RMSE: **0.8942**
- Minimum k for which MAE converges: **16,** with average MAE: **0.6815**

Question 6

- **Popular Movie Trimming:**
    - Minimum average RMSE: **0.8728**

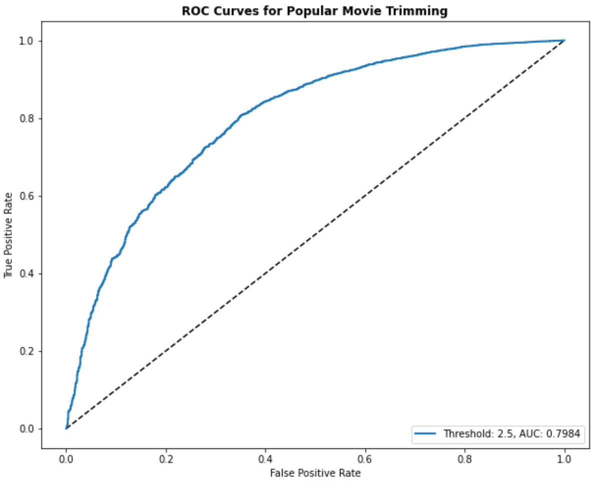
k-NN collaborative filter with 10-fold cross validation on Popular Movie Trimming

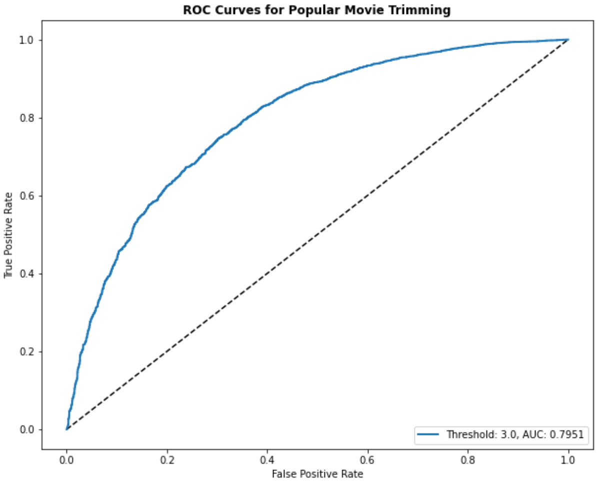


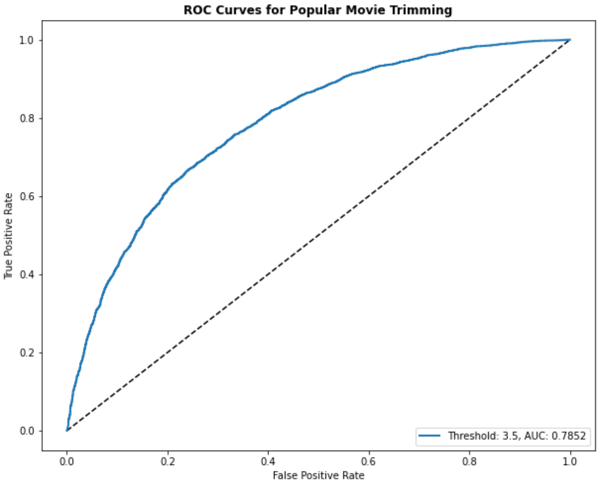
- ROC Curve for different threshold:
    - Because the RMSE pattern is similar to the pattern of question 4 and 5, We calculate minimum k value for popular movie training testset to plot ROC curves. We choose **k = 16**.

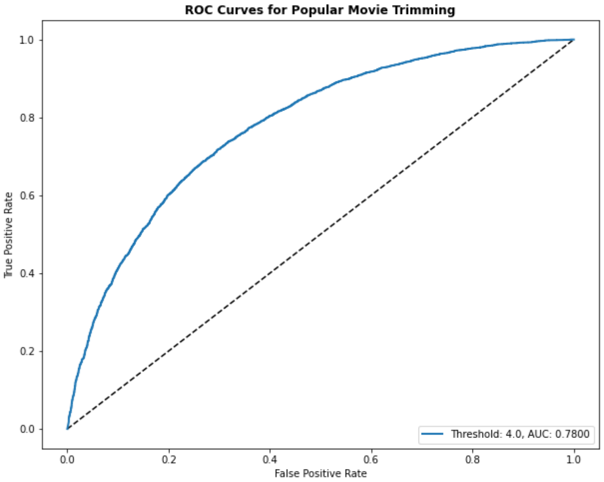
Threshold	AUC value
2.5	0.7984

3.0	0.7951
3.5	0.7852
4.0	0.7800

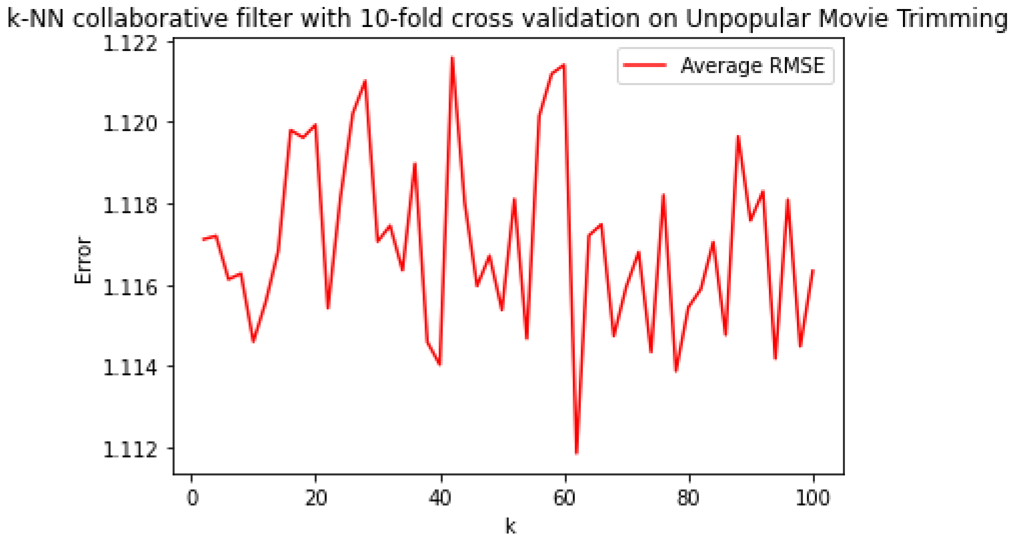






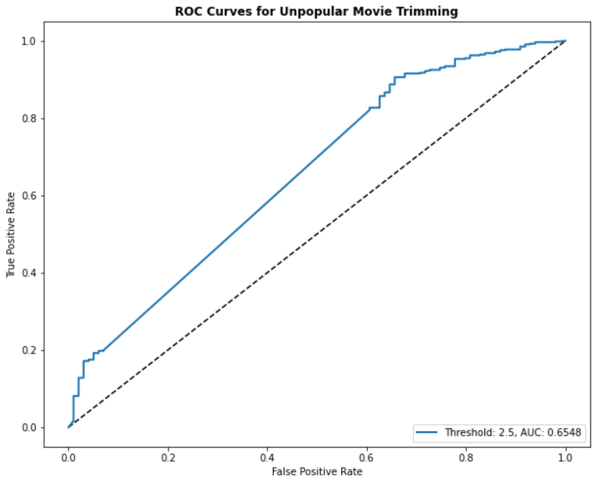


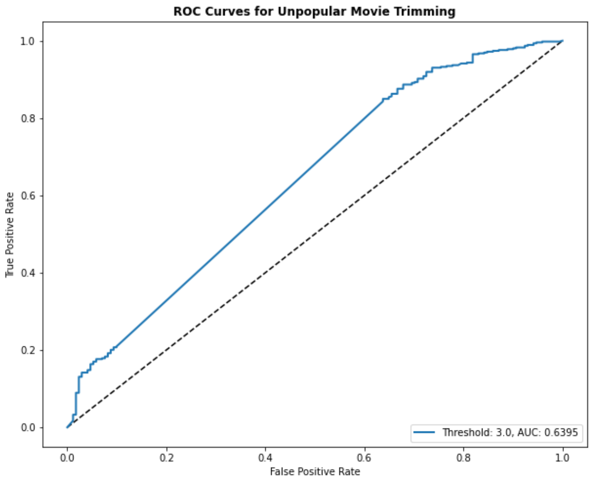
- **Unpopular Movie Trimming**:
    - Minimum average RMSE: **1.1119**

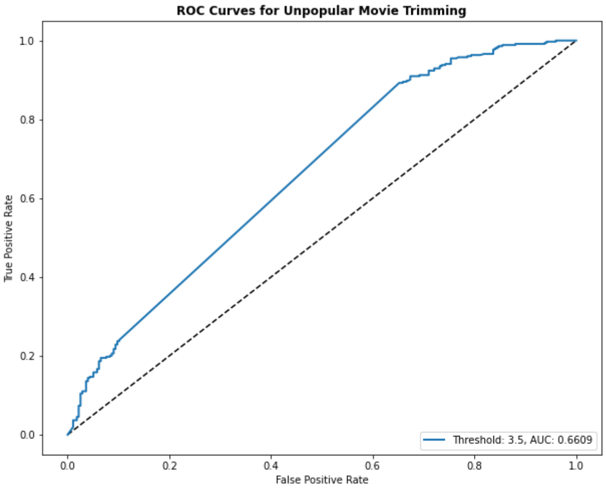


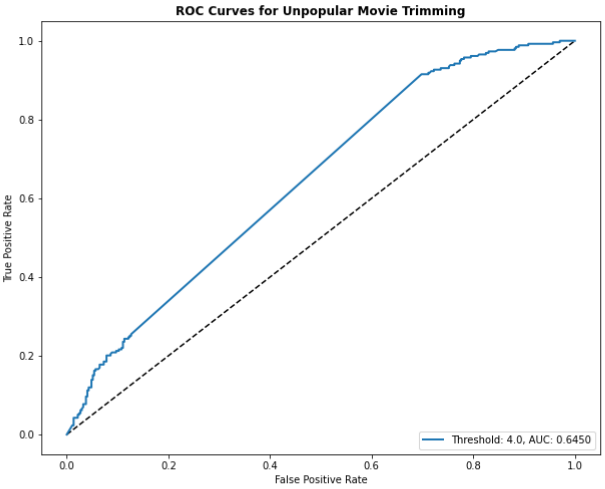
- ROC Curve for different threshold:
    - We eyeball the k value with the least average RMSE error: **k = 62**.

Threshold	**AUC value**
2.5	0.6548
3.0	0.6395
3.5	0.6609
4.0	0.6450

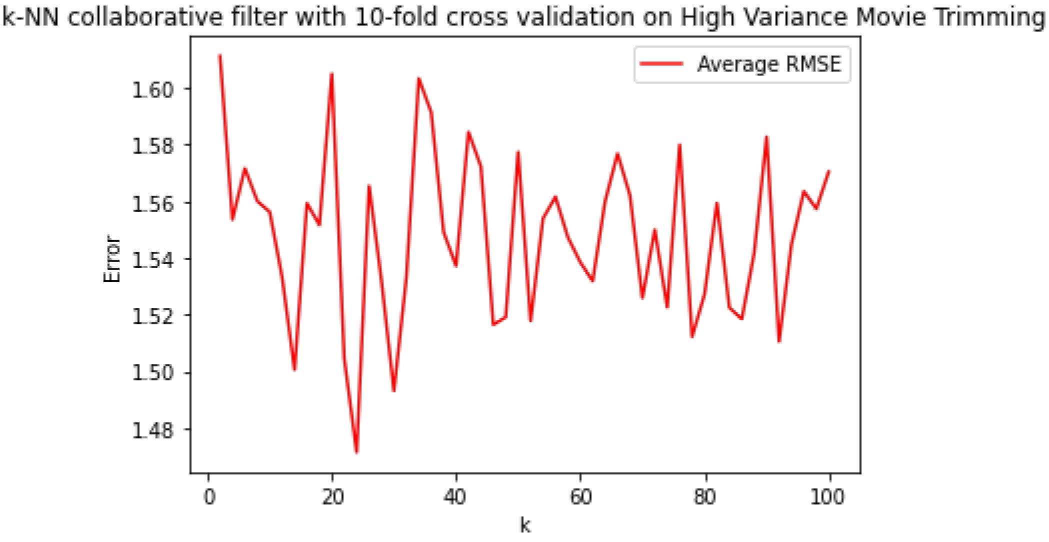






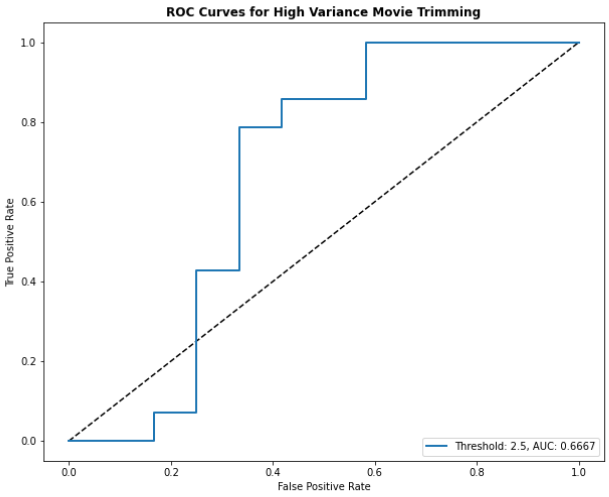


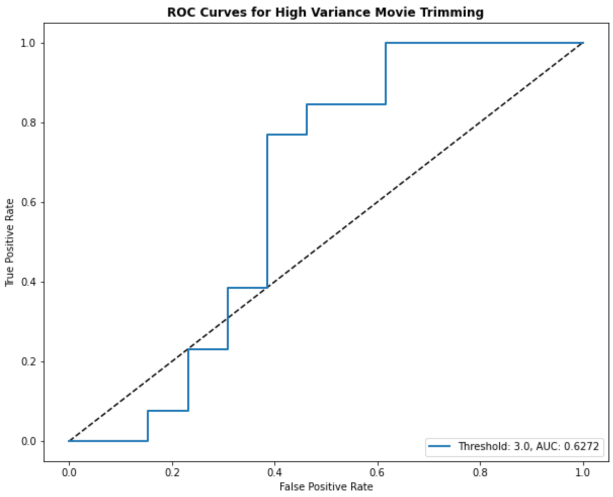
- **High Variance Movie Trimming:**
    - Minimum average RMSE: **1.4715**

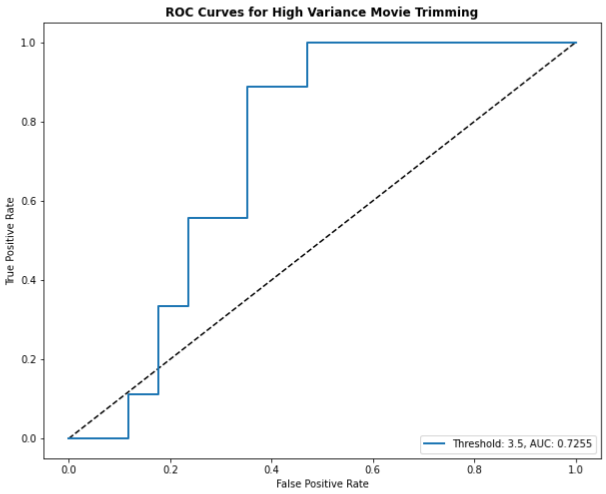


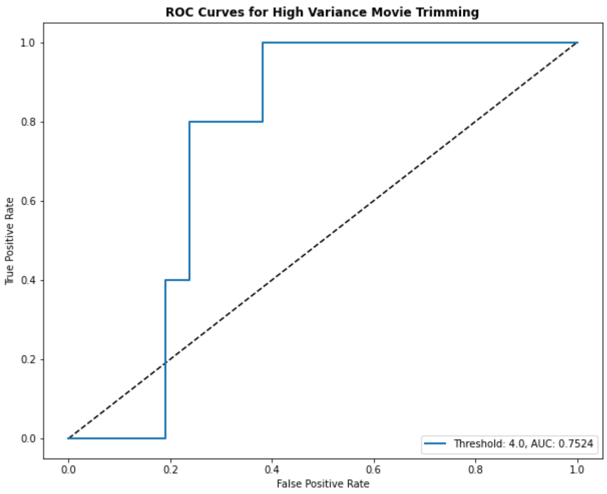
- ROC Curve for different threshold:
    - We eyeball the k value with the least average RMSE error: **k = 24**.

Threshold	AUC value
2.5	0.6667
3.0	0.6272
3.5	0.7255
4.0	0.7524









Question 7

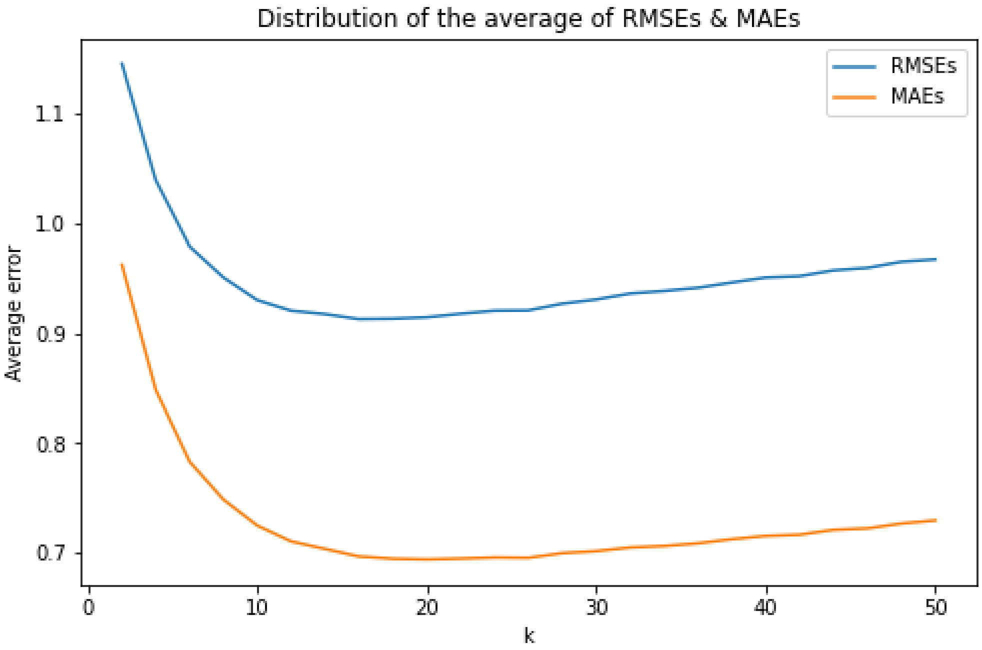
- The optimization problem given by equation 5 is certainly **not** convex. If U is fixed, then the least-square problem will be as follows,

$$min_V \sum_{i=1}^{m}\sum_{j=1}^{n} W_{ij}(r_{ij} \ - \ (UV^T)_{ij})^2$$

- Because of the matrix W, the minimization problem cannot be convex.

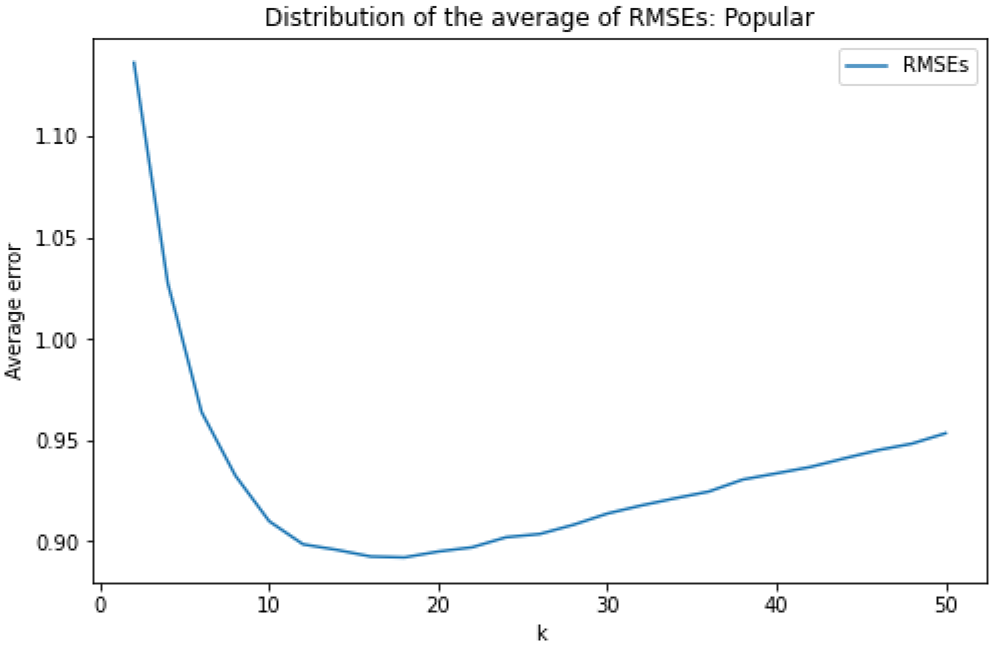
Question 8

- A



- B
    - RMSE: optimal number of latents = **18**, MIN RMSE = **0.9137**
    - MAE: optimal number of latents = **18**, MIN MAE = **0.6947**
    - Total number of genres = **19**

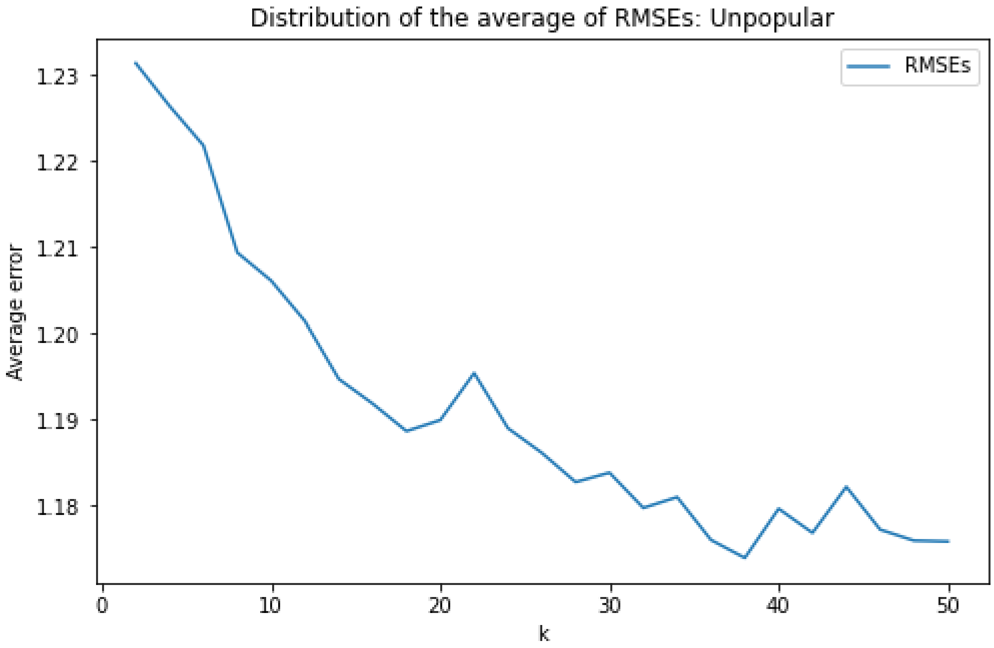
- Based on the results from part a and the plot, we found that the optimal number of latents is approximately the same as the number of movie genres.
- C
    - **Popular Trimming**



- Popular: optimal number of latents = 18, **MIN RMSE = 0.8920**

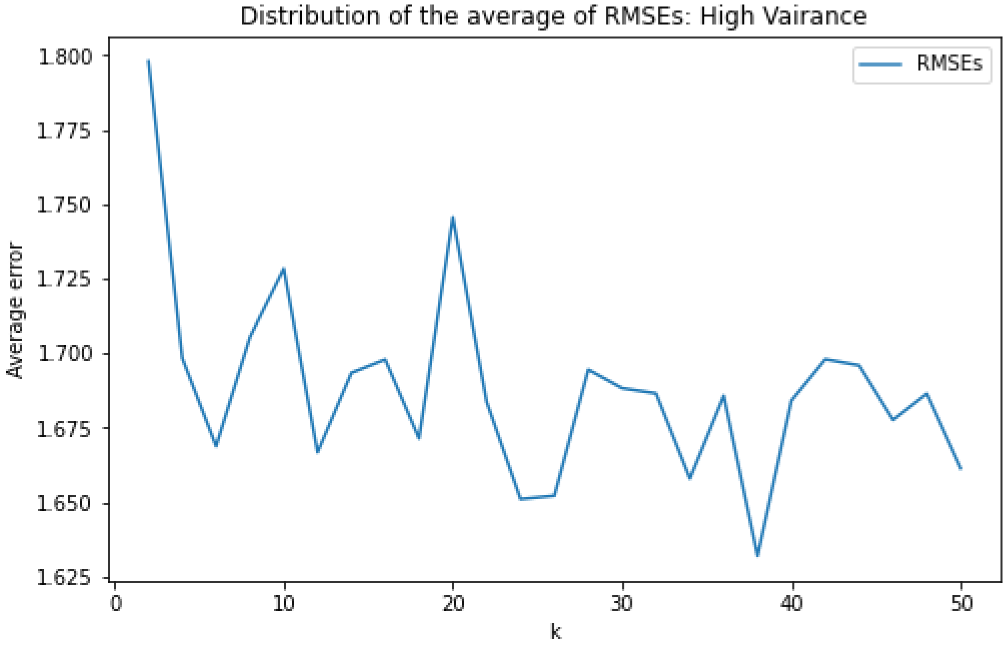
        - ROC curves:
            - We use n_factors = 18

- **Unpopular Trimming**



    - High Variance: optimal number of latents = 38, **MIN RMSE = 1.1739**

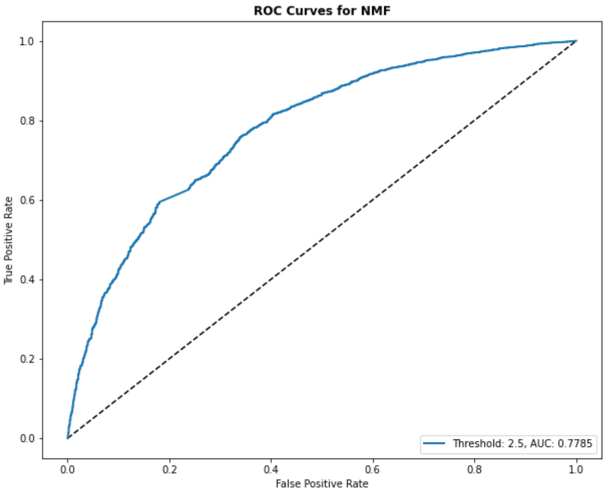
- **High Variance Trimming**

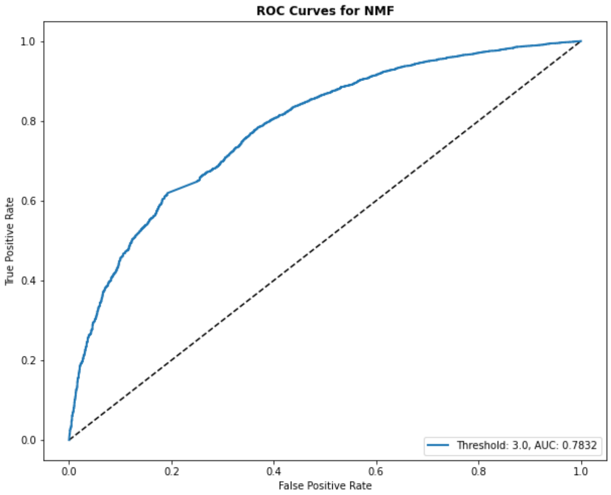


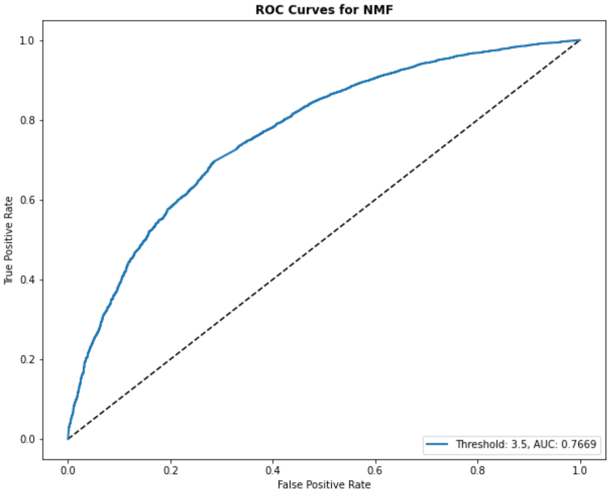
    - High Variance: optimal number of latents = 38, **MIN RMSE = 1.6320**

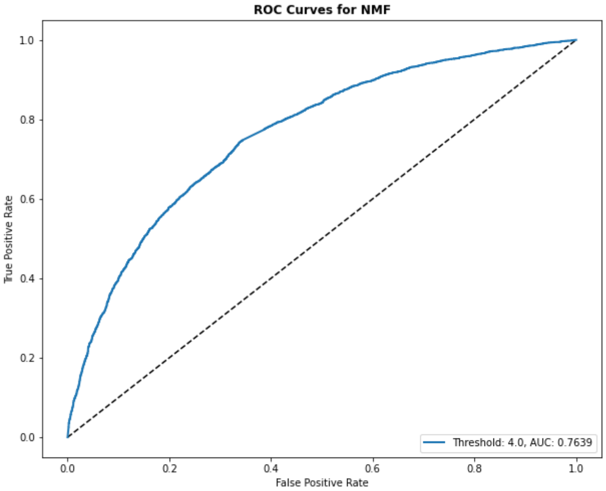
- ROC

Threshold	AUC value
2.5	0.7785
3.0	0.7832
3.5	0.7669
4.0	0.7639









Question 9

- In this question, we choose 3 representatives of latent factors to analyze:
  ****

  Latent Factor 0
  Genre: Children|Fantasy|Musical
  Genre: Comedy
  Genre: Drama|Romance|War
  Genre: Drama|Romance

Genre: Drama|Romance
Genre: Horror|Mystery|Thriller
Genre: Comedy|Drama|Romance
Genre: Adventure|Drama|Sci-Fi
Genre: Comedy|Drama
Genre: Comedy|Romance
****

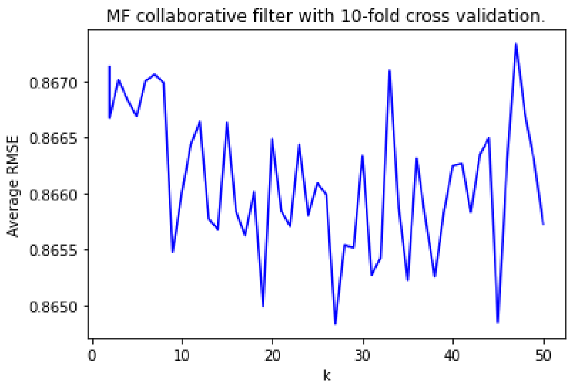
Latent Factor 7
Genre: Action|Drama
Genre: Action|Drama|Thriller
Genre: Adventure|Drama|Romance
Genre: Comedy
Genre: Drama
Genre: Drama
Genre: Action|Crime|Thriller|IMAX
Genre: Drama
Genre: Animation|Children|Fantasy|IMAX
Genre: Sci-Fi
****

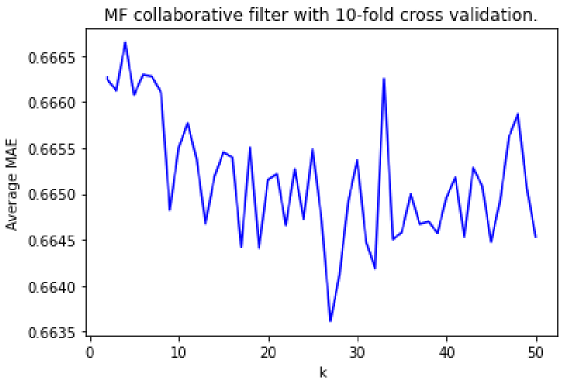
Latent Factor 16
Genre: Animation|Children|Comedy
Genre: Documentary
Genre: Comedy|Drama|Romance
Genre: Comedy|Drama|Romance
Genre: Comedy
Genre: Action|Adventure|Crime|Thriller
Genre: Adventure|Children|Comedy
Genre: Documentary
Genre: Adventure|Horror|Sci-Fi
Genre: Adventure|Animation|Comedy|Fantasy|Romance|Sci-Fi
****

- We can see that **each later factor represents different genre of movies**:
    - For latent factor 0, most of the movies are drama/romance movies
    - For latent factor 6, most of the movies are action/thriller movies
    - For latent factor 16, most of the movies are comedy/animation movies
- When the number of latent factors increases, the number of distinct movie genres decreases.
- The movie genres show a better clustering when the number of latent factors increases.

Question 10

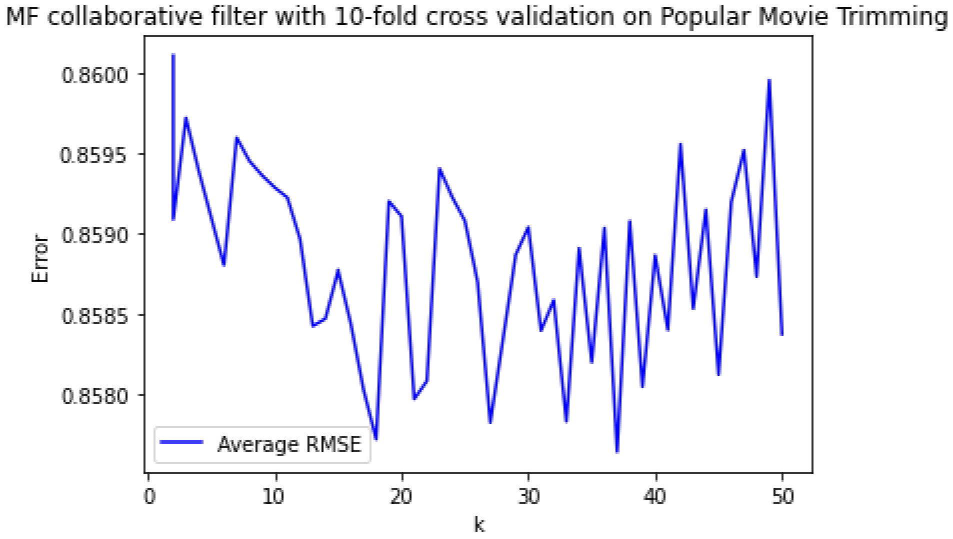
- A





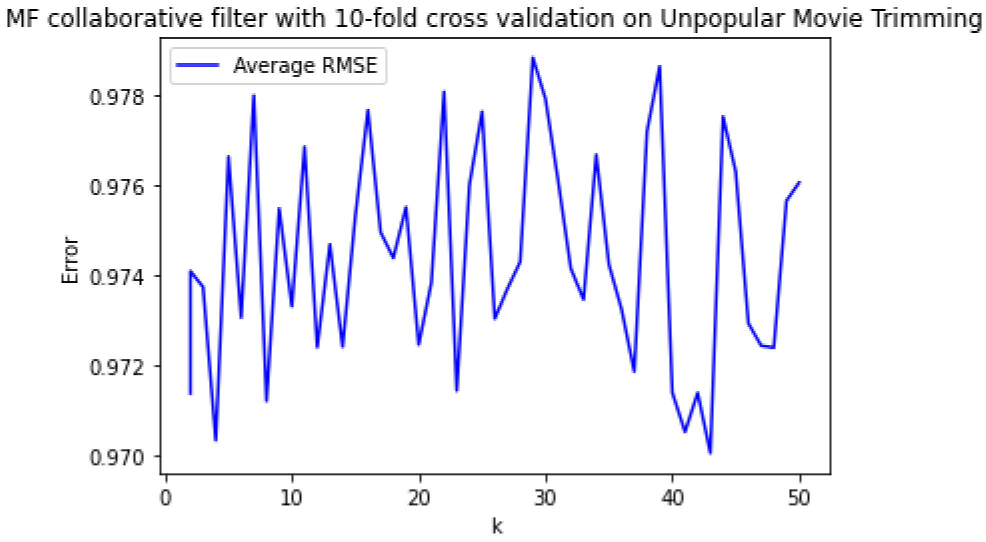
- B
    - Minimum average RMSE: 0.8648; Optimal number of latent factors: **27**
    - Minimum average MAE: 0.6636; Optimal number of latent factors: **27**
    - The optimal number of latent factors 27 is **not close to** the total number of movie genres 19. Hence latent factors in MF model are **not very interpretable.**

- C
    - **Popular Trimming**



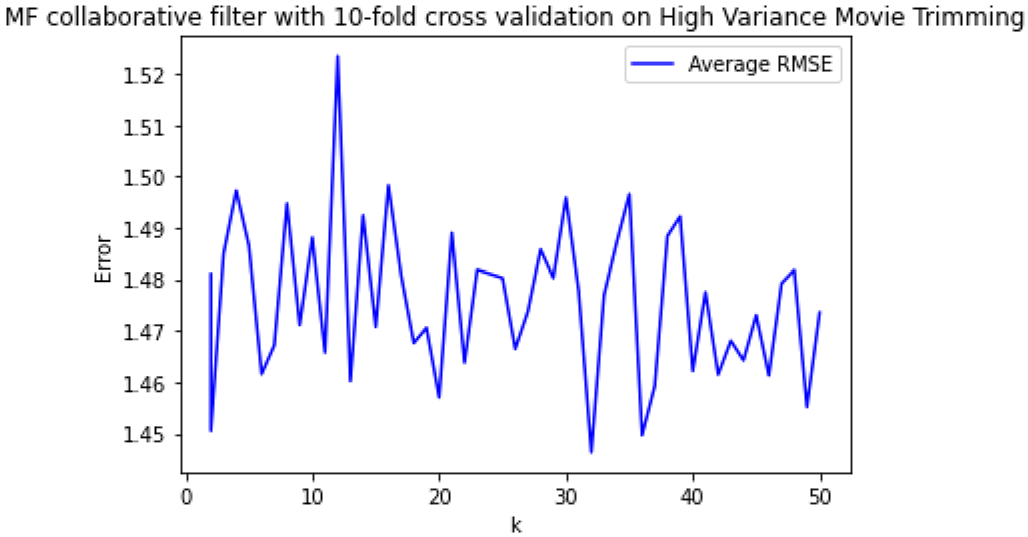
- Minimum average RMSE for popular movie trimming dataset: **0.8576**

- **Unpopular Trimming**



- Minimum average RMSE for unpopular movie trimming dataset: **0.9700**

- **High Variance Trimming**

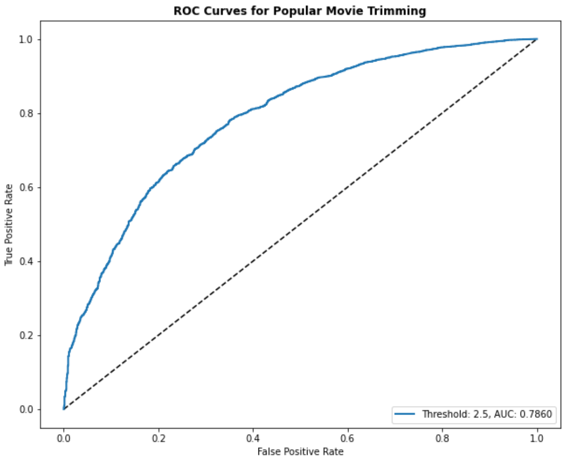


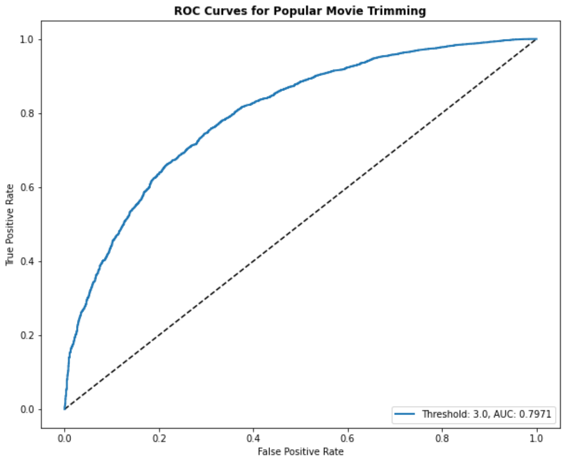
- Minimum average RMSE for high variance movie trimming dataset: **1.4463**

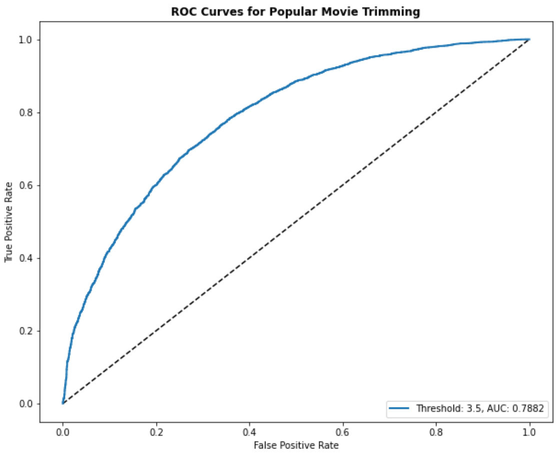
- ROC

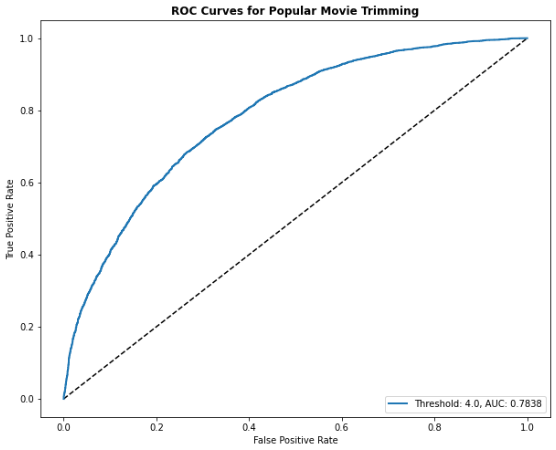
Threshold	AUC value
2.5	0.7860

3.0	0.7971
3.5	0.7882
4.0	0.7838







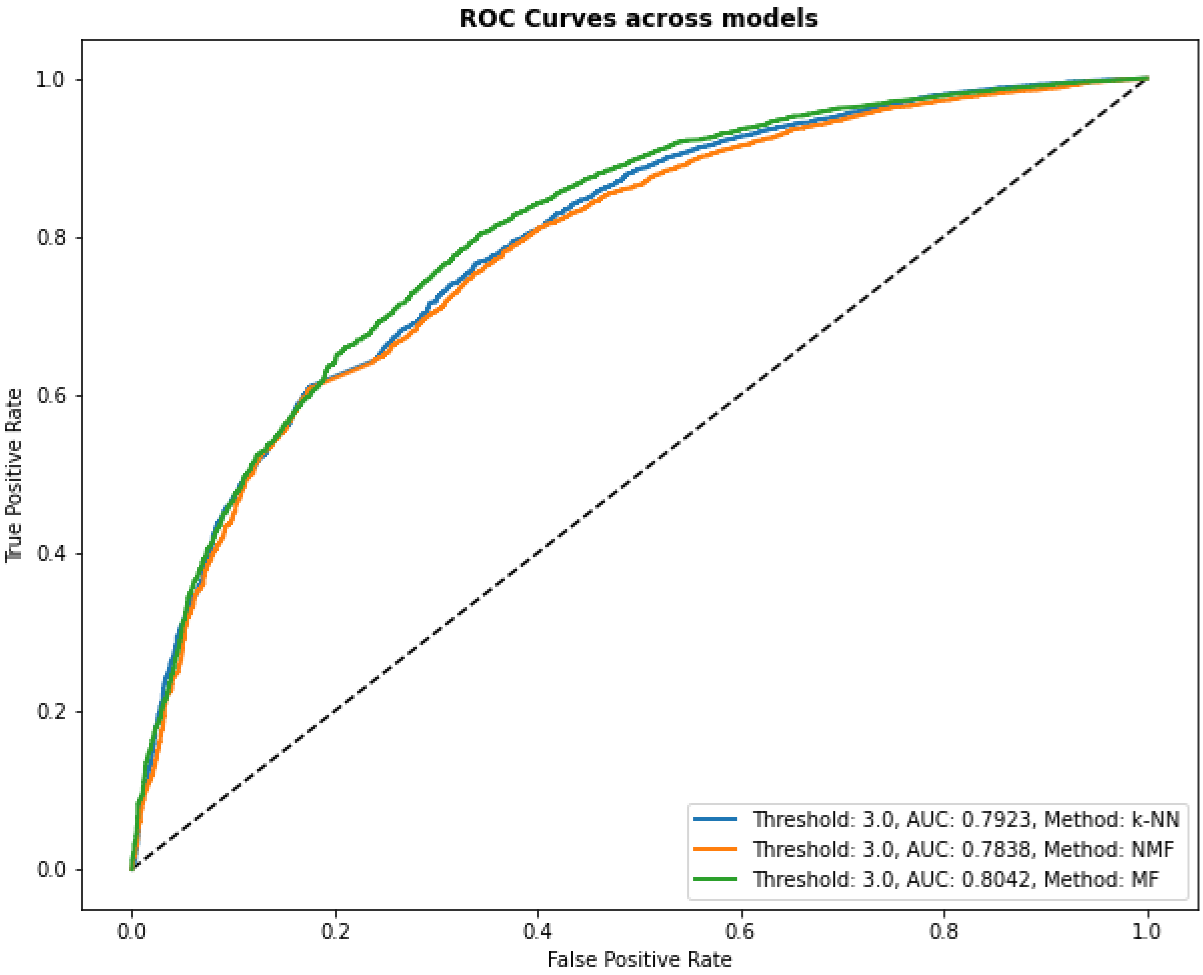


Question 11

- For the naive collaborative filter, we can simply extract the all ratings given by each user, and then calculate the mean ratings given by each user, and form a dictionary.
- Performance on the original dataset
    - Average RMSE original dataset: **0.9347**
- Performance on Test set subsets
    - Average RMSE Popular: **0.9323**
    - Average RMSE Unpopular: **0.9711**
    - Average RMSE High-Variance: **1.4348**

Question 12

- Use k = 16 for k-NN, k = 18 for NMF, and k = 27 for MF.



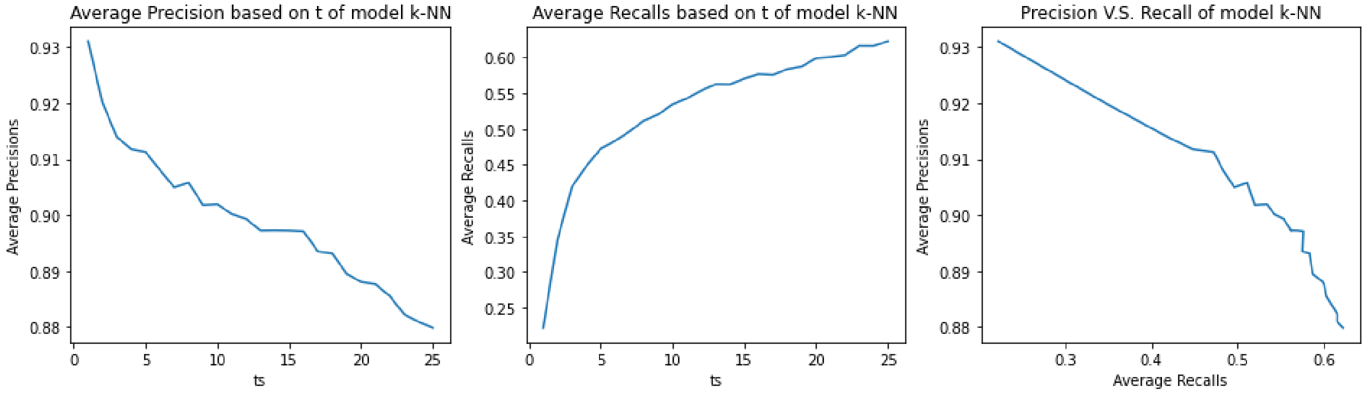
- We can use the area under the curve to measure the performance. Compared to the three different models, **MF with bias collaborative filtering model** reaches the highest AUC score. This indicates that MF with bias is the best for predicting the ratings of the movies. Compared to the NNMF collaborative filtering model, MF with bias collaborative filtering model has optimization variables for bias on both users and movies side, which is more flexible. This nature helps MF models to drop the extreme data in which users give abnormally high or low ratings and reduce the bias introduced by users.

Question 13

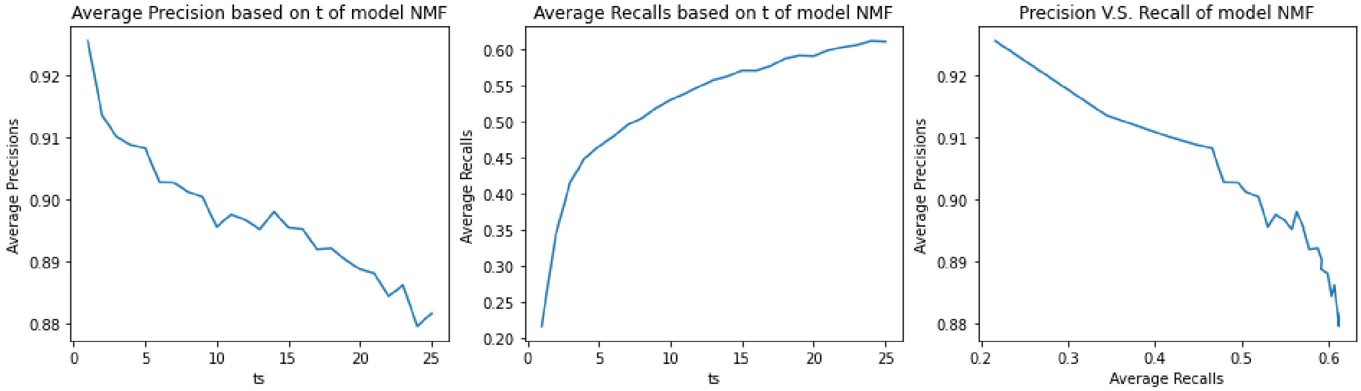
- **Precision**
  - Based on our datasets, precision is the fraction of user-**liked** items out of the items recommended by the system to the user. Namely, it reflects how relevant the prediction of the recommendation system is.
- **Recall**
  - Based on our datasets, recall is the fraction of the items that are **recommended** to the user and user **likes**, out of **all items that a user likes**. In other words, it tells how many true values are predicted by the recommendation system.

Question 14

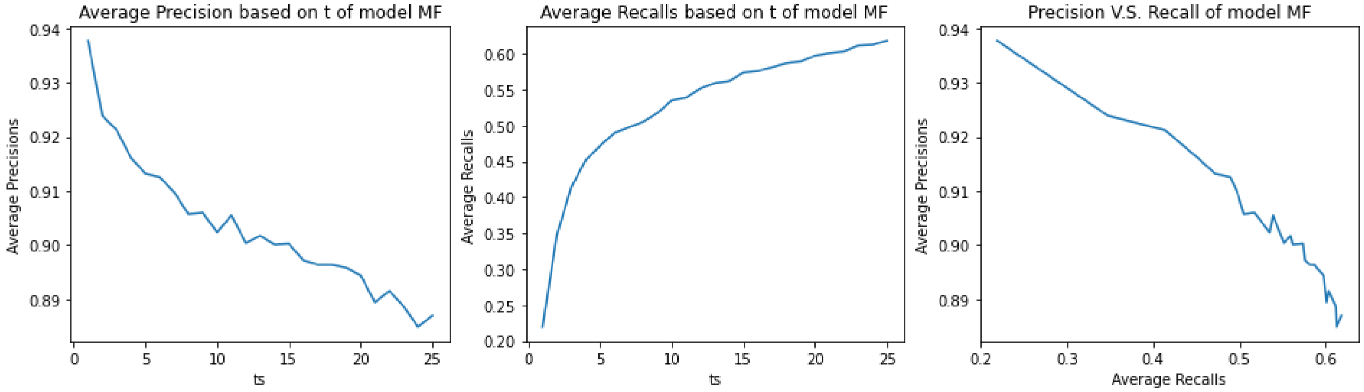
- k-NN



- NMF



- MF



- Similar to AUC in ROC curve, area under precision vs recall curve (plots in third column) can be treated as the measurement to compare the three recommendation systems. From the plots we can observe biased MF achieves the best performance, while NMF has the worst performance. Besides, we can also observe an inverse relationship between precision and recall. The first column of the plots shows precisions and ts are inversely related, while the second column shows recalls and ts are positively correlated. Therefore, given a fix ts, the precisions and recalls will be inversely related, as shown in the third column.