Project 3: Recommendation System

EECS219 Winter 2023

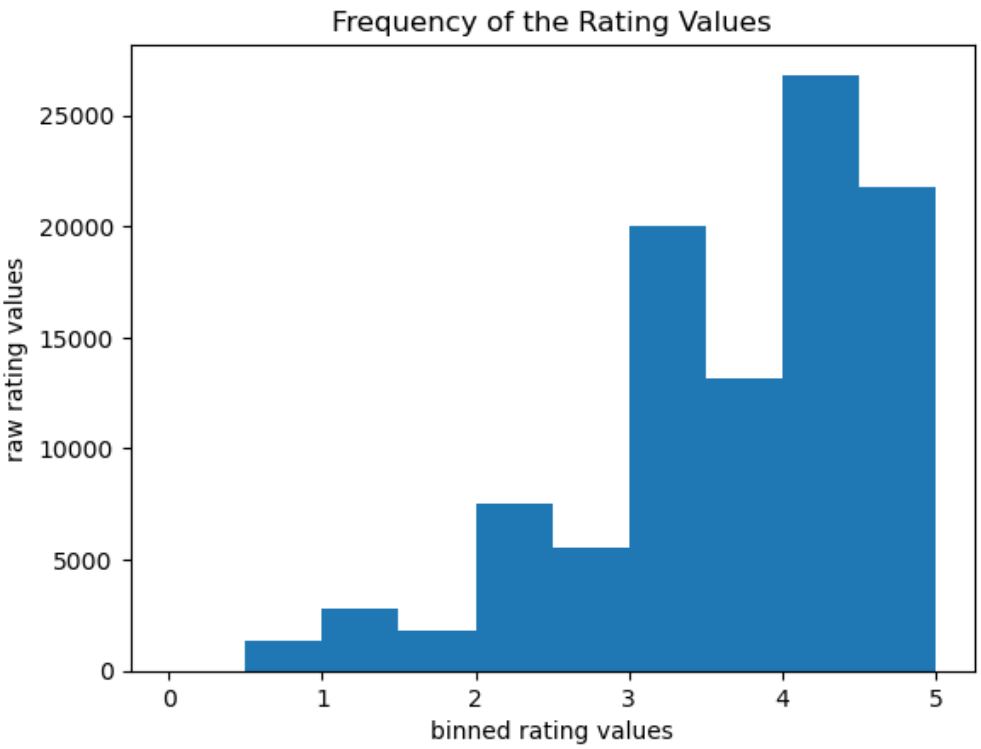
Fengyuan Heying, Alice Lu, Dadian Zhu

Question 1:

- A:

Sparsity:	0.016999683055613623

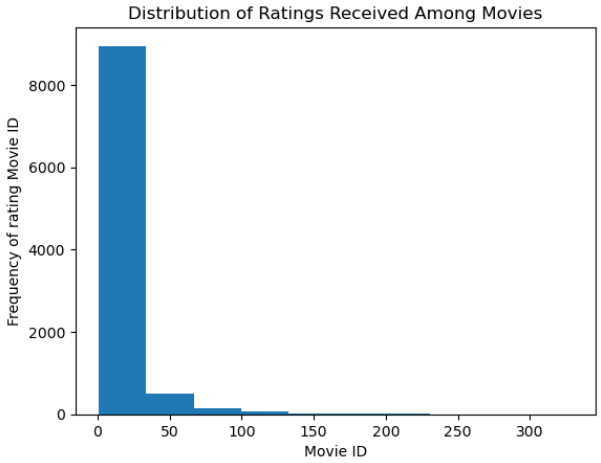
- B:



Frequency of the Rating Values

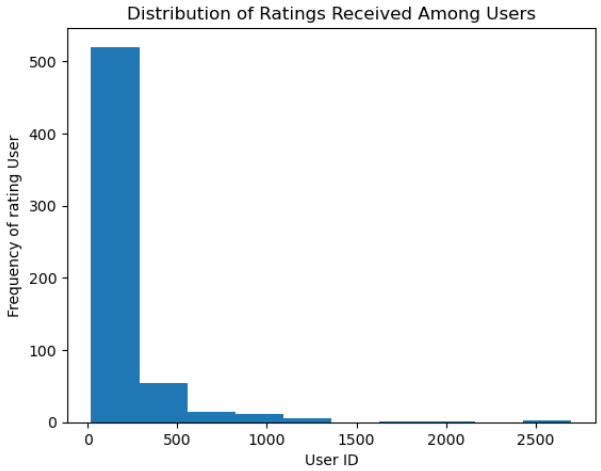
Shape of the raw rating values approximately has a trend of increasing but non-monotonically.

- C:



Distribution of Ratings Received among Movies

- D:

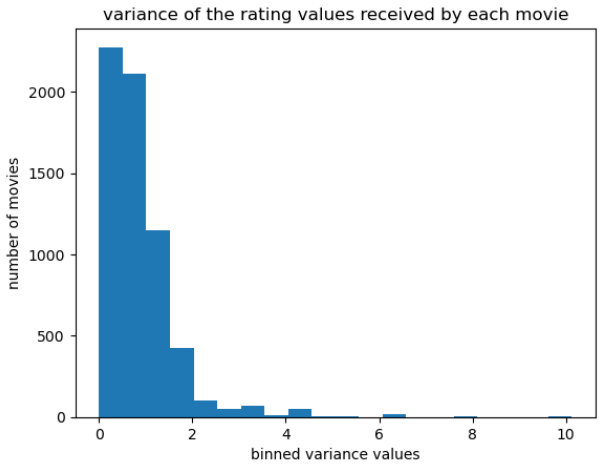


Distribution of Ratings Received among Users

- E:

  Distribution of both graphs has a decreasing trend. Both of them have high frequency low movie ID and user ID. It shows that lower movie ID was rated and people has low user ID rate more movies than high user ID.

- F:



Variance of the Rating Values Received by Each Movie

The shape of the variance plot has a decreasing trend. Most of the variance values focus on intervals 0 to 2. After variance value=2, there's not too many movies.

Question 2:

- A:

$$\mu_u = \frac{\sum_{i \in I_u} r_{ui}}{||I_u||}$$

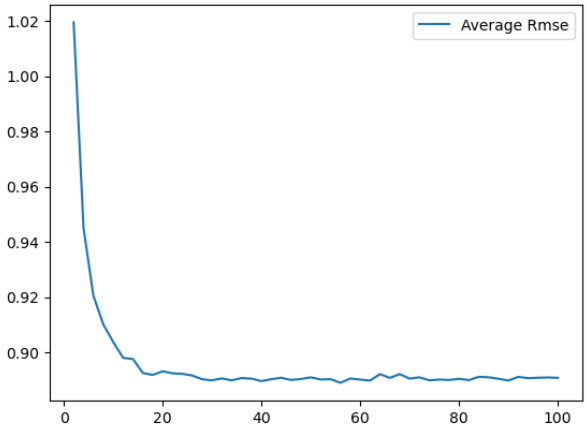
- B:

It means user u and user v have the same set of rating item indices. It can be an empty set because u and v may not rate the same movie. Also rating matrix R is sparse and also shows that it is possible for users u and v to rate totally different movies.

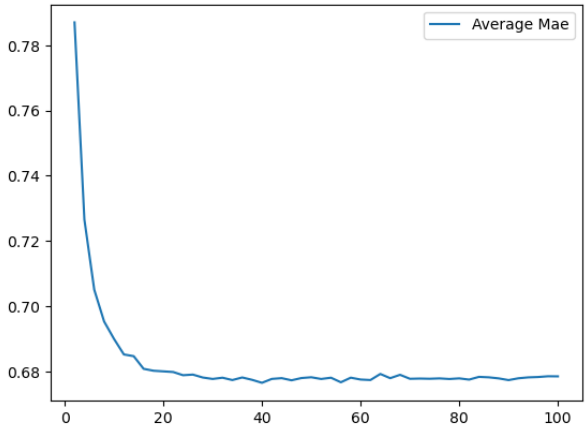
Question 3:

Mean-centering the raw ratings redefine the point for the predictor to be centered. It shifts the scale over, but retains the units. The effect is that the slope between the predictor and rating value we predict does not change at all. For instance, if there are users A and B, and both of them have extreme rating habits. A only rated the movie from 1 to 3, and B only rated the movie 4 to 6. If we don't subtract the mean, we will predict a similar rating for A and B which is not accurate.

Question 4:



Average RMSE against k

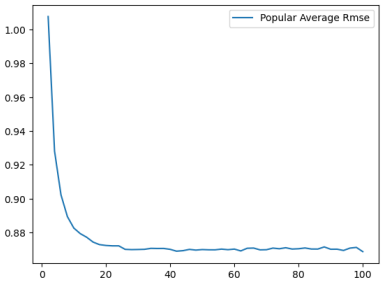


Average MAE against k

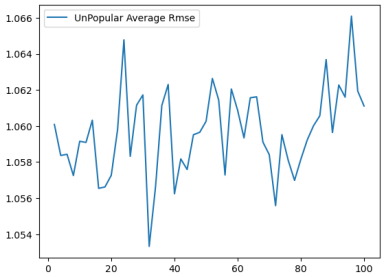
Question 5:

k value:	12

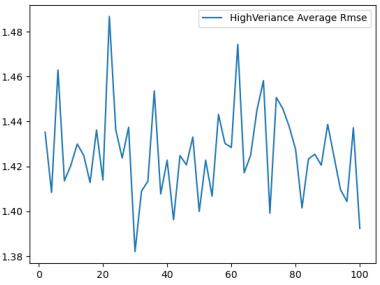
Question 6:



Popular Average RMSE

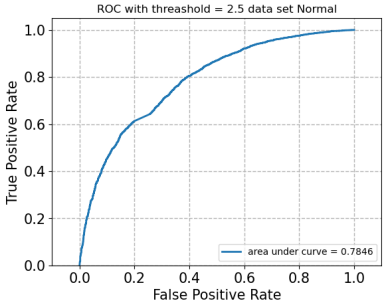


Unpopular Average RMSE

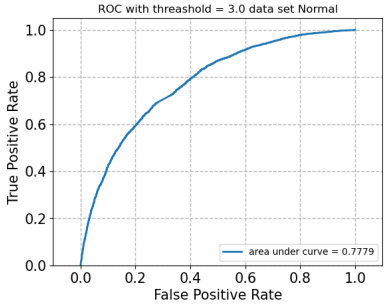


High-Variance Average RMSE

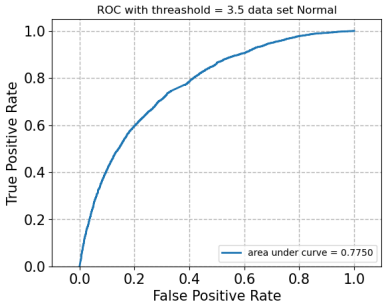
ROC with different threshold set at Normal



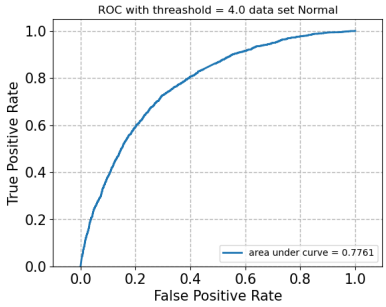
threshold=2.5



threshold=3.0

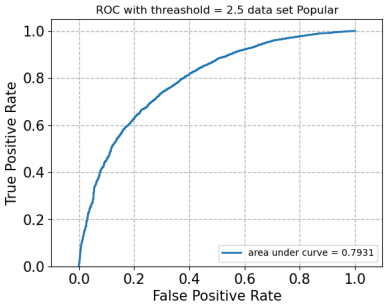


threshold=3.5

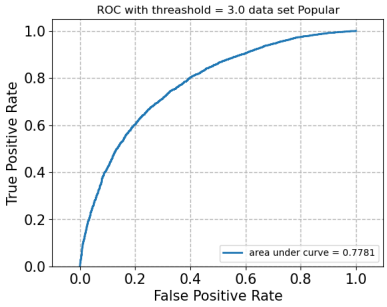


threshold=4.0

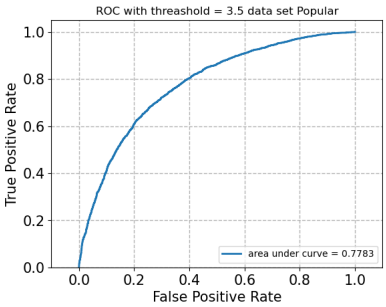
ROC with different threshold set at Popular



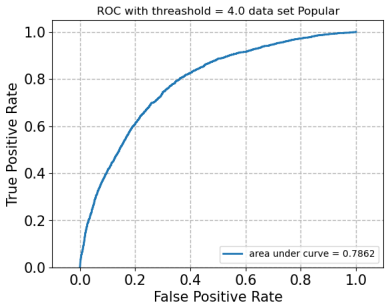
threshold=2.5



threshold=3.0

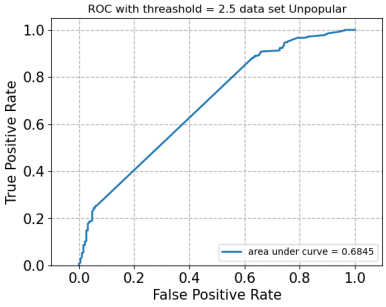


threshold=3.5

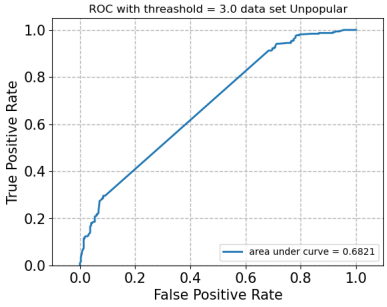


threshold=4.0

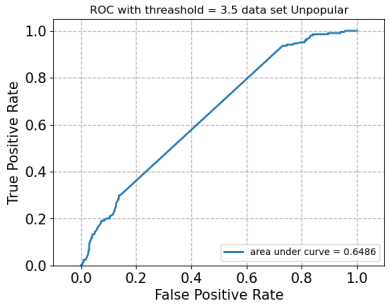
ROC with different threshold set at Unpopular



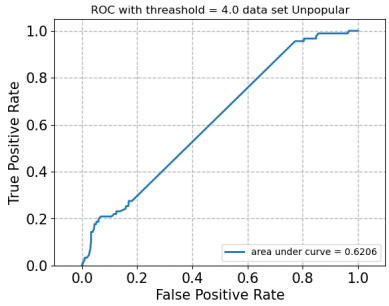
threshold=2.5



threshold=3.0

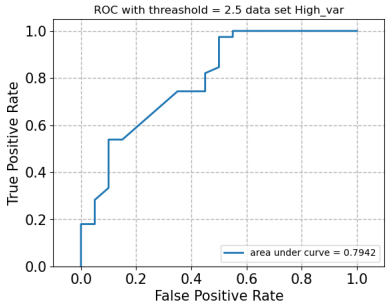


threshold=3.5

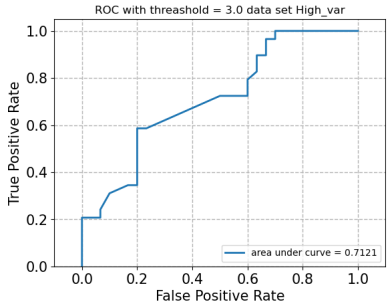


threshold=4.0

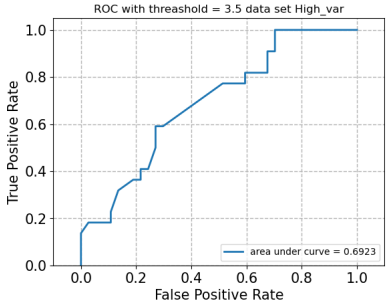
ROC with different threshold set at High Variance



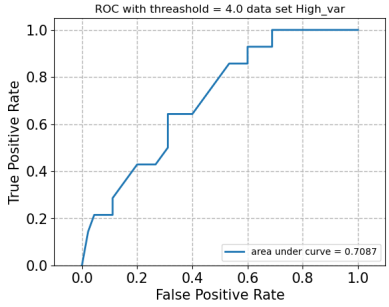
threshold=2.5



threshold=3.0



threshold=3.5



threshold=4.0

Question 7:

Yes, with fixed U, the equation is just as a least square problem where r is y and U is k and V^T is x. Least square is convex thus this equation is convex.

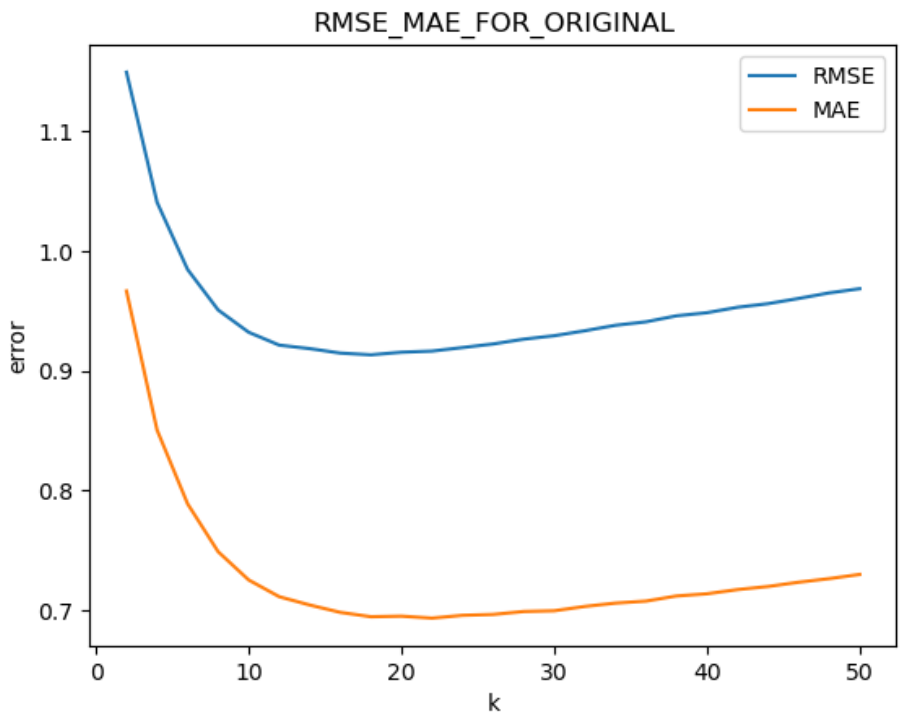
Question 8:

- A:

```
k_dict=train_and_predict(data_ratings)
```

```
Done with k of value 2, get an avgRMSE of 1.1489562171988301 and an avgMAE of
0.9664932586450607
Done with k of value 4, get an avgRMSE of 1.040512317383016 and an avgMAE of
0.8505933039063329
Done with k of value 6, get an avgRMSE of 0.98443444801166 and an avgMAE of
0.7890123449539304
Done with k of value 8, get an avgRMSE of 0.9506656090717446 and an avgMAE of
0.7490546442678965
Done with k of value 10, get an avgRMSE of 0.9320661606579481 and an avgMAE o
f 0.7254818372221062
Done with k of value 12, get an avgRMSE of 0.9213370943644241 and an avgMAE o
f 0.7115198380622083
Done with k of value 14, get an avgRMSE of 0.9184265015572961 and an avgMAE o
f 0.7045920957943321
Done with k of value 16, get an avgRMSE of 0.9147151389639345 and an avgMAE o
f 0.6984552026112235
Done with k of value 18, get an avgRMSE of 0.9132728166746364 and an avgMAE o
f 0.6948384242116434
Done with k of value 20, get an avgRMSE of 0.9154437320989057 and an avgMAE o
f 0.6953094095703346
Done with k of value 22, get an avgRMSE of 0.9163727817954028 and an avgMAE o
f 0.6936572816831418
Done with k of value 24, get an avgRMSE of 0.9193353200841035 and an avgMAE o
f 0.6960090131247041
Done with k of value 26, get an avgRMSE of 0.9223144692592076 and an avgMAE o
f 0.6967272322860973
Done with k of value 28, get an avgRMSE of 0.9263207952299908 and an avgMAE o
f 0.6991962577935411
Done with k of value 30, get an avgRMSE of 0.9291879618328271 and an avgMAE o
f 0.6997944654135742
Done with k of value 32, get an avgRMSE of 0.9333287074236413 and an avgMAE o
f 0.7033981701539871
Done with k of value 34, get an avgRMSE of 0.9378808883950864 and an avgMAE o
f 0.7061851639700014
Done with k of value 36, get an avgRMSE of 0.9407730640608438 and an avgMAE o
f 0.7078471713877379
Done with k of value 38, get an avgRMSE of 0.9457877466351615 and an avgMAE o
f 0.7121276791354977
Done with k of value 40, get an avgRMSE of 0.9483889065154377 and an avgMAE o
f 0.7139591226106226
Done with k of value 42, get an avgRMSE of 0.9529282384248224 and an avgMAE o
f 0.717468596832344
Done with k of value 44, get an avgRMSE of 0.955935705323745 and an avgMAE of
0.7200532222785825
Done with k of value 46, get an avgRMSE of 0.9602693847304966 and an avgMAE o
f 0.723605234251925
Done with k of value 48, get an avgRMSE of 0.964914372654035 and an avgMAE of
0.7265945283455006
Done with k of value 50, get an avgRMSE of 0.9683942623475799 and an avgMAE o
f 0.7301098284817259
```

List of average RMSE and MAE across all 10 folds

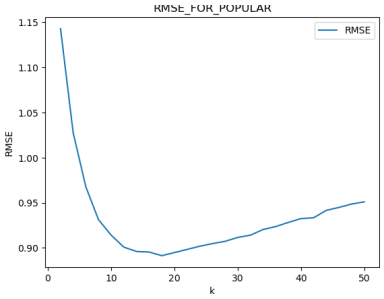


Average RMSE and MAE against k

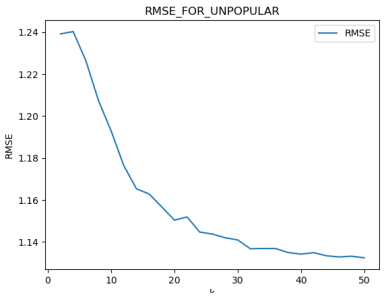
- B:

  The minimum average RMSE is 0.9132728166746364 and minimum average MAE is 0.6948384242116434. This happens when we have a k of 18 which is the same as the number of movie genres.

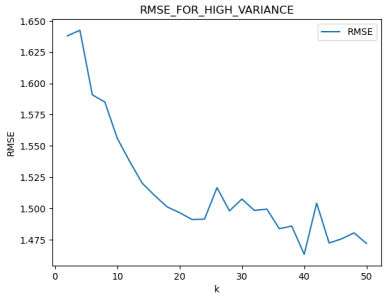
- C:



Popular Average RMSE

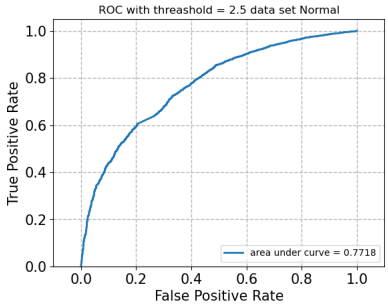


Unpopular Average RMSE

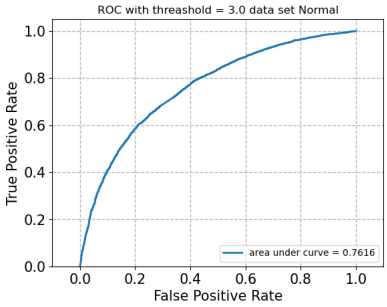


High-Variance Average RMSE

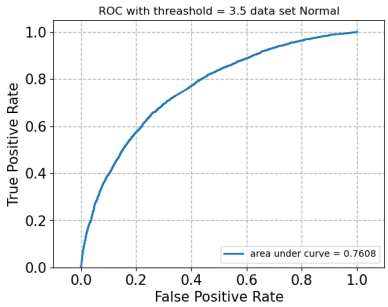
ROC with different threshold set at Normal



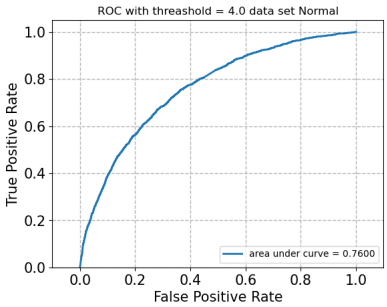
threshold=2.5



threshold=3.0

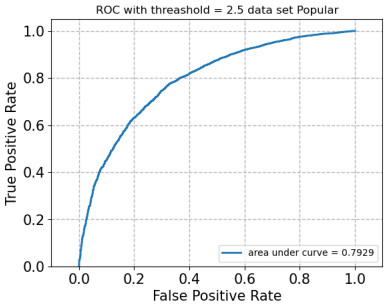


threshold=3.5

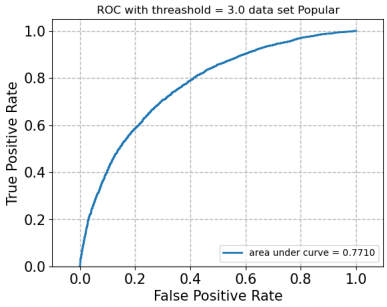


threshold=4.0

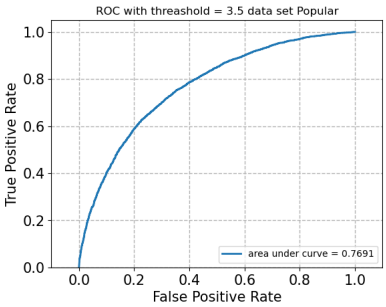
ROC with different threshold set at Popular



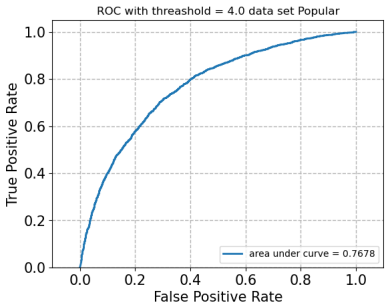
threshold=2.5



threshold=3.0

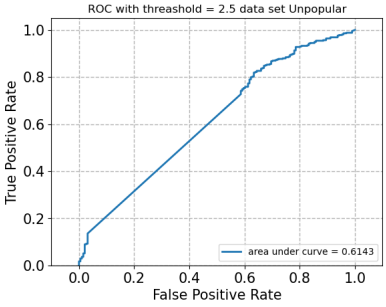


threshold=3.5

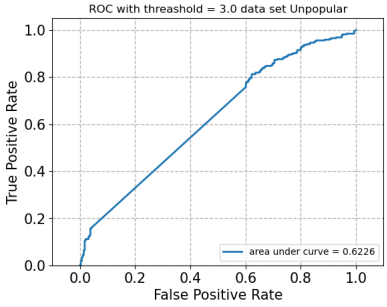


threshold=4.0

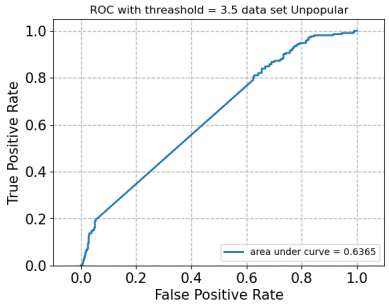
ROC with different threshold set at Unpopular



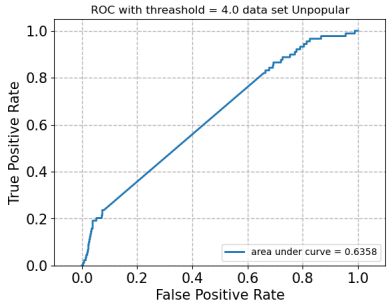
threshold=2.5



threshold=3.0

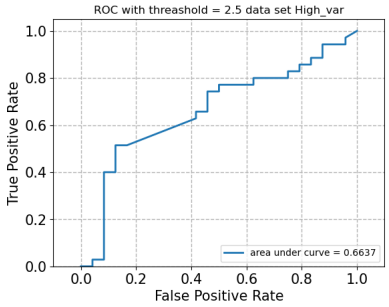


threshold=3.5

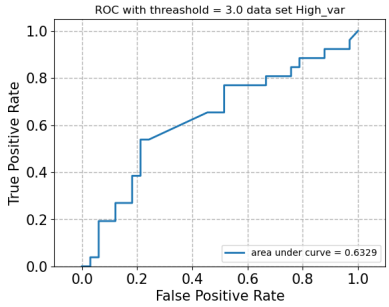


threshold=4.0

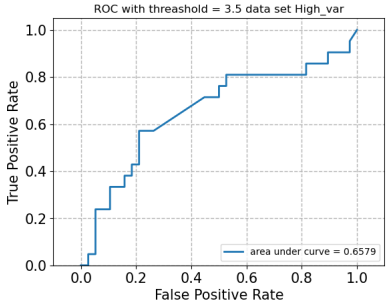
ROC with different threshold set at High Variance



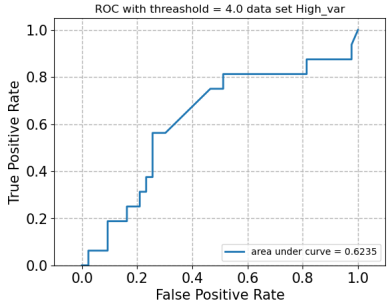
threshold=2.5



threshold=3.0



threshold=3.5



threshold=4.0

Question 9:

```
[83]: for col in all_cols_top_10_genres:
          print(col)
```

```
['Fantasy|Horror', 'Horror|Mystery|Thriller', 'Horror', 'Action|Crime|Thrille
r', 'Drama|Thriller', 'Drama|Horror|Thriller', 'Drama', 'Comedy|Drama', 'Acti
on|Adventure', 'Crime|Horror|Mystery']
['Children|Comedy', 'Comedy', 'Comedy|Fantasy|Horror', 'Comedy|Romance', 'Adv
enture|Comedy|Sci-Fi', 'Drama|War', 'Crime|Drama', 'Action|Comedy|Sci-Fi|Thri
ller', 'Comedy', 'Action|Adventure|Children|Comedy']
['Comedy|Drama|Romance', 'Comedy|Crime', 'Action|Sci-Fi|Thriller', 'Documenta
ry', 'Documentary', 'Romance|Thriller', 'Action|Children', 'Comedy', 'Comedy|
Romance', 'Drama|War']
['Action|Sci-Fi', 'Sci-Fi', 'Horror|Sci-Fi', 'Comedy|Drama', 'Comedy', 'Actio
n|Drama|Fantasy', 'Horror|Sci-Fi', 'Action|Adventure|Animation|Children|Comed
y|Romance', 'Drama', 'Drama|Romance|Western']
['Action|Fantasy|Horror|Sci-Fi|Thriller', 'Comedy|Romance', 'Drama', 'Drama|H
orror|Thriller', 'Action|Drama', 'Comedy|Fantasy', 'Thriller', 'Horror|Sci-Fi
|Thriller', 'Comedy|Documentary', 'Comedy|Crime']
['Animation|Children|Comedy|Musical', 'Adventure|Animation|Children|Comedy|IM
AX', 'Horror|Thriller', 'Comedy|Romance', 'Animation|Children|Comedy', 'Horro
r', 'Action|Comedy|Crime', 'Comedy|Drama|Romance', 'Action|Comedy', 'Comedy|D
rama']
['Children|Comedy|Fantasy|Romance', 'Action|Adventure|Animation', 'Action|Com
edy|Crime|Fantasy', 'Comedy|Drama', 'Children|Comedy|Drama', 'Comedy', 'Actio
n|Comedy|Crime|Thriller', 'Drama', 'Action|Comedy', 'Horror|Mystery']
['Action|Adventure|Sci-Fi', 'Drama|Romance', 'Documentary', 'Action|Drama',
'Animation|Comedy', 'Drama|Mystery', 'Comedy|Fantasy|Horror|Musical|Thrille
```

Genres for few columns

```
In [85]: sortedFreq

Out[85]: [('Comedy', 87),
          ('Drama', 81),
          ('Action', 41),
          ('Thriller', 36),
          ('Romance', 35),
          ('Horror', 33),
          ('Sci-Fi', 27),
          ('Fantasy', 22),
          ('Adventure', 19),
          ('Crime', 17),
          ('Children', 15),
          ('Mystery', 13),
          ('Animation', 10),
          ('Documentary', 6),
          ('War', 4),
          ('Western', 3),
          ('Musical', 3),
          ('IMAX', 2),
          ('Film-Noir', 1)]
```

Top 10 movies

Most top 10 movies belong to Comedy.

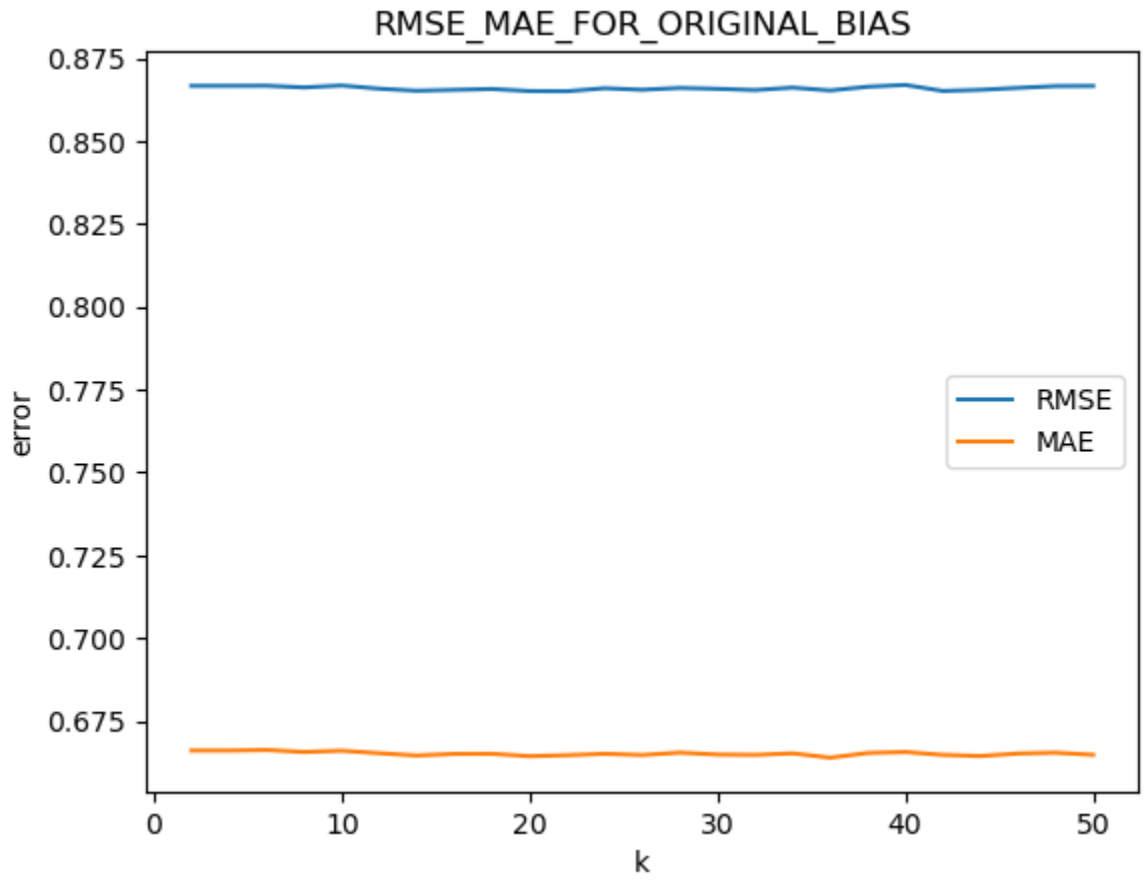
Connection between latent factor and movie genres: each row i in the item latent factor indicates factors for the ith item in the original dataset. The ith item in the original dataset has a corresponding moveID. We can use the corresponding movieID to fetch the genre of the movie using the movies.csv dataset.

Question 10:

- A:

Minimum RMSE	0.8651113157471088
Minimum MAE	0.6638930756707422
Optimal k with Best RMSE	22

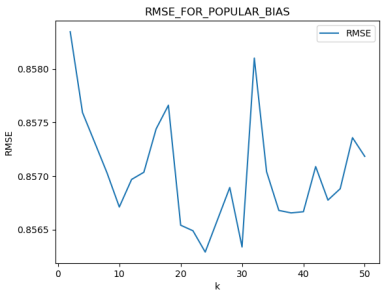
- B:



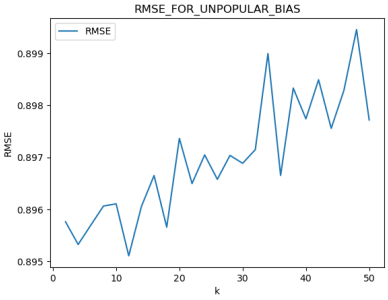
RMSE and MAE for original bias

As we can see from the figure, the RMSE and MAE do not change significantly as we change k. Besides, the minimum RMSE we get is when k is 22. Therefore the optimal k is 22, which is not the same as the number of movie genres.

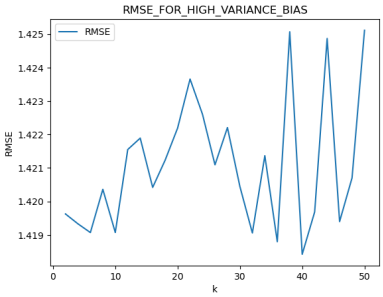
- C:



Popular Bias RMSE

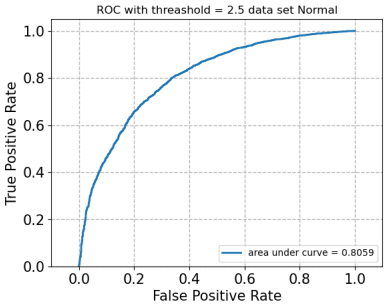


Unpopular Bias RMSE

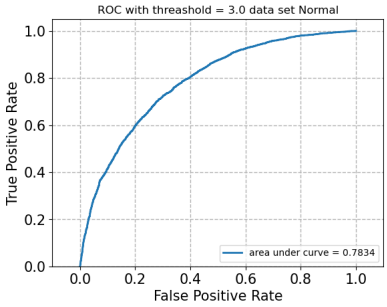


High-Variance Bias RMSE

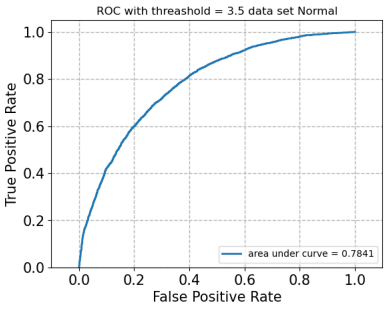
ROC with different threshold set at Normal



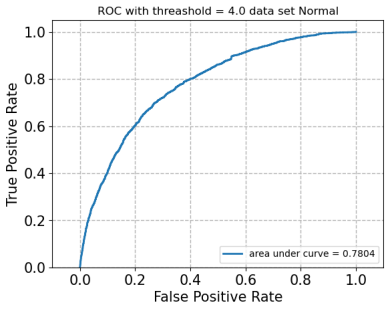
threshold=2.5



threshold=3.0

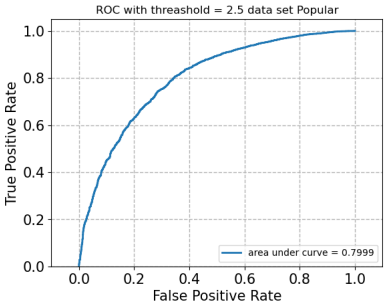


threshold=3.5

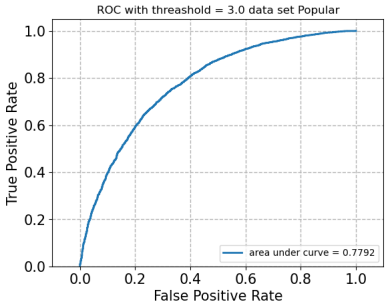


threshold=4.0

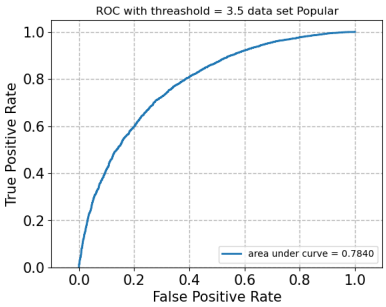
ROC with different threshold set at Popular



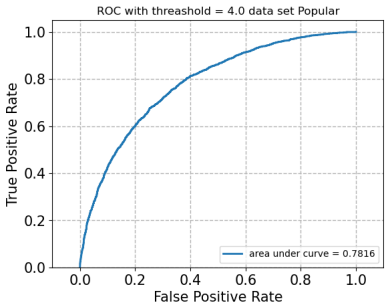
threshold=2.5



threshold=3.0

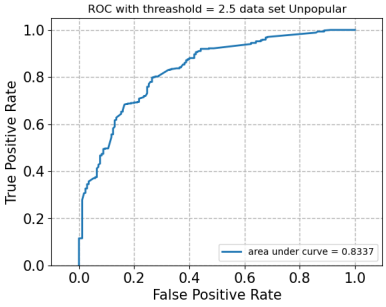


threshold=3.5

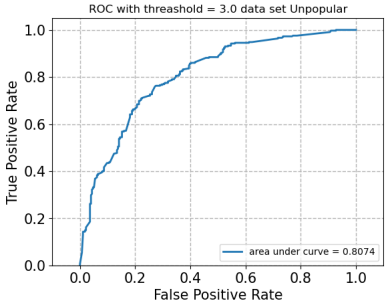


threshold=4.0

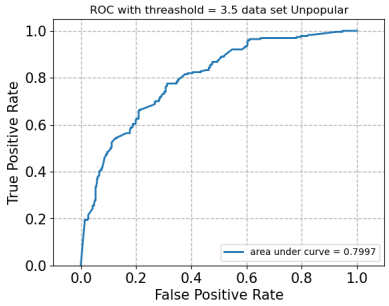
ROC with different threshold set at Unpopular



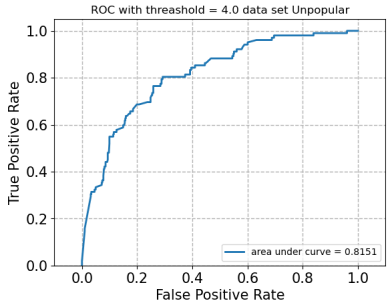
threshold=2.5



threshold=3.0

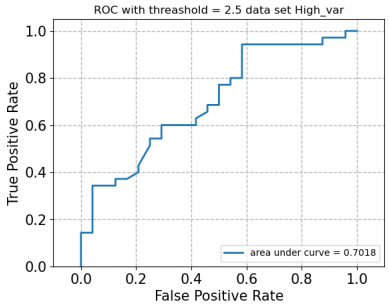


threshold=3.5

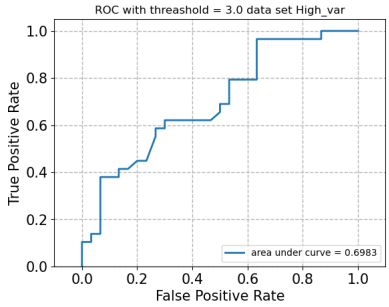


threshold=4.0

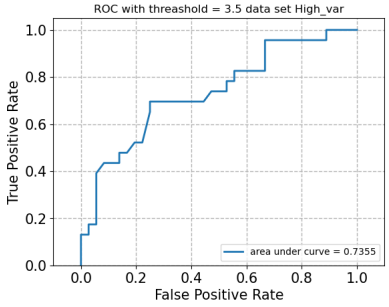
ROC with different threshold set at High Variance



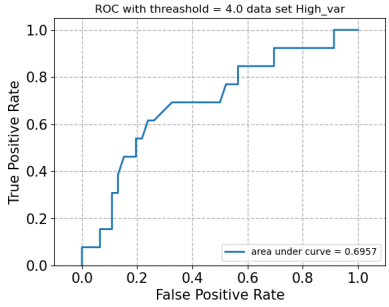
threshold=2.5



threshold=3.0



threshold=3.5



threshold=4.0

Question 11:

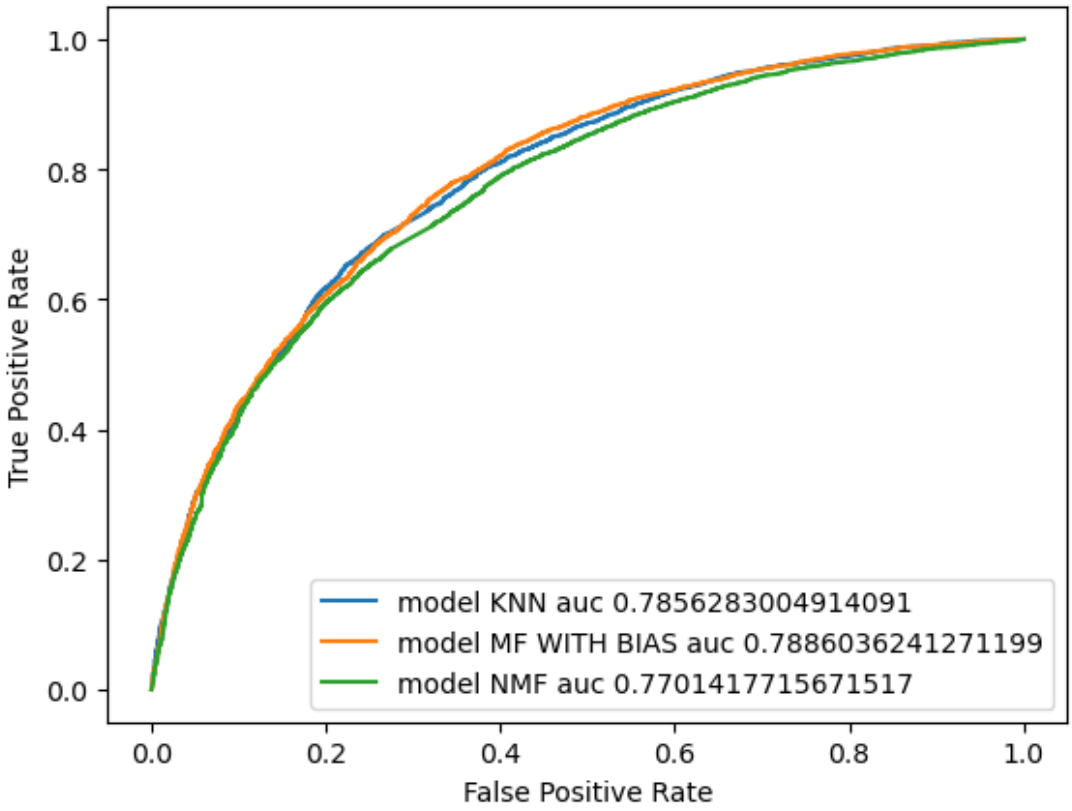
- A:

average RMSE for original dataset:	0.9410603557679573

- B:

Data subset	Average RMSE
Popular	0.937438851087791
Unpopular	1.0092265456916887
High Variance	1.9810700212393992

Question 12:



Best ROC for different models

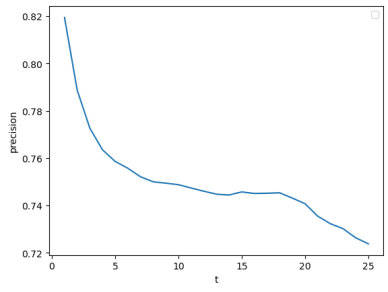
The AUC here indicates the performance of the collaborative filters by telling how the collaborative filters distinguish ratings above threshold and ratings below threshold. As the figure shows, AUC of MF with bias is the largest. Thus, performance of MF with bias is the best among these three models. KNN performs just a little bit worse than MF with bias. NMF performs the worst among these three models.

Question 13:

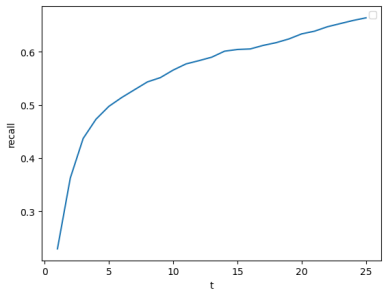
precision	among all items we recommended to the user, how many items are truly liked by the user.
recall	among all items the user likes, how many items that user actually like are recommended to the user by us.

Question 14:

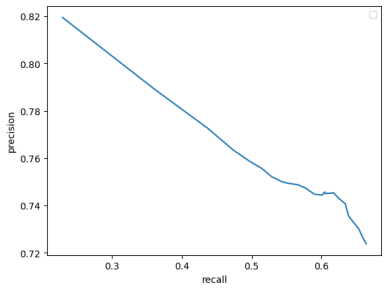
NMF Models



Average Precision against t



Average Recall against t



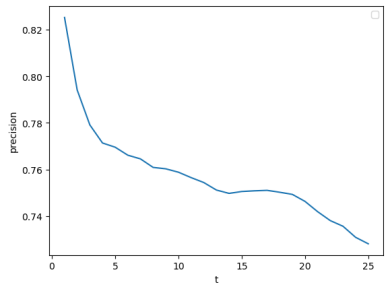
Best k recall curves

Shape of plot NMF:

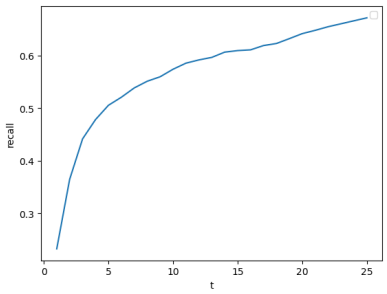
- Precision of the model decreases as we increase it, it is like the shape of an exponential graph.
- Recall of the model increases as we increase t, it is like the shape of an exponential graph.

- Precision of the model decreases as the recall of the model increases, it is like the shape of a linear graph.

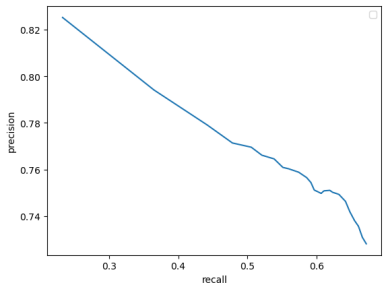
KNN Models



Average Precision against t



Average Recall against t

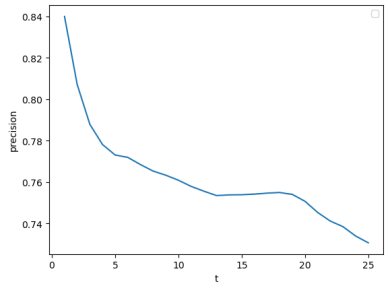


Best k recall curves

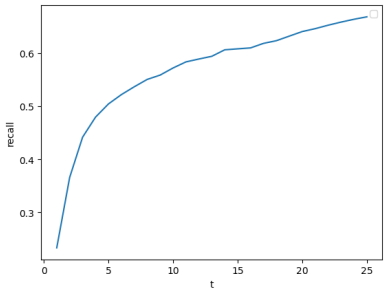
Shape of plot KNN:

- Precision of the model decreases as we increase it, it is like the shape of an exponential graph.
- Recall of the model increases as we increase it, it is like the shape of an exponential graph.
- Precision of the model decreases as the recall of the model increases, it is like the shape of a linear graph.

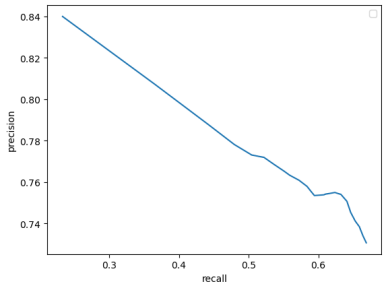
MF Models



Average Precision against t



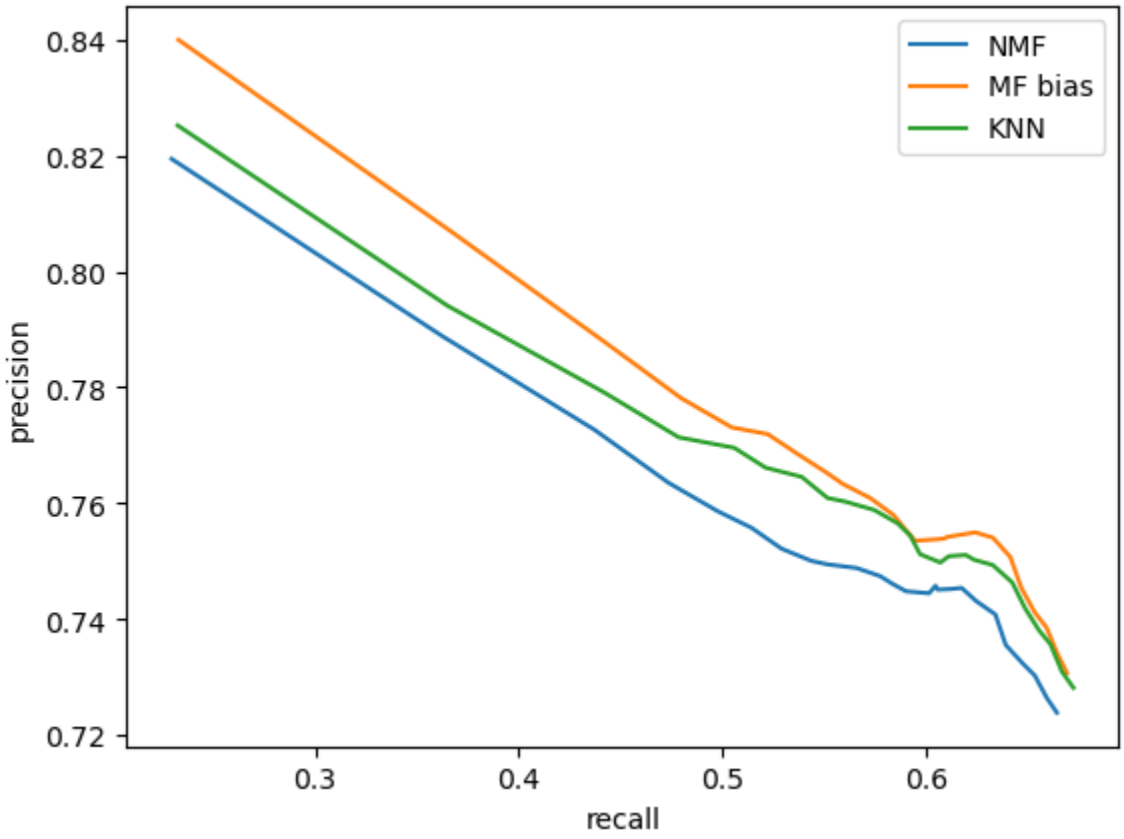
Average Recall against t



Best k recall curves

Shape of plot MF with bias:

- Precision of the model decreases as we increase it, it is like the shape of an exponential graph.
- Recall of the model increases as we increase it, it is like the shape of an exponential graph.
- Precision of the model decreases as the recall of the model increases, it is like the shape of a linear graph



Comparison of three models in precision-recall curves

As the figure shows, among movies MF bias recommends, most recommended movies are liked by users as we can see from precision and also among movies user likes, MF recommend most of them as we can see from recall. Even if the recall for the model increases, the precision of the

model is still good. Thus, MF bias performs the best among these three models KNN performs a little bit worse than MF with bias, NMF performs the worst among these three models