# **Group Members:**

XINGQUAN GUAN, UID: 804 759 254

DONG WANG, UID: 804 363 328

PENG-YU CHEN, UID: 704 732 316

ZHENGXIANG YI, UID: 304 761 570

# **Acknowledgement:**

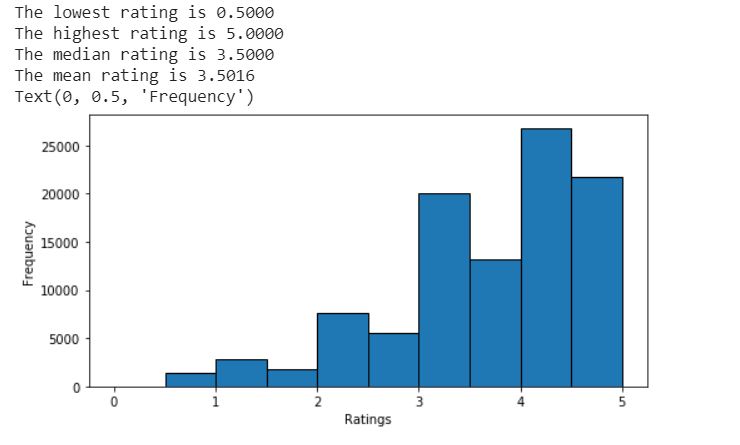
All group members **have made equal contributions** to the project and same grading scale is expected for individual.

We accomplished this project using *Google Colab*. If anyone wants to test the submitted codes, please run it in *Google Colab* environment.

## **Question 1.**

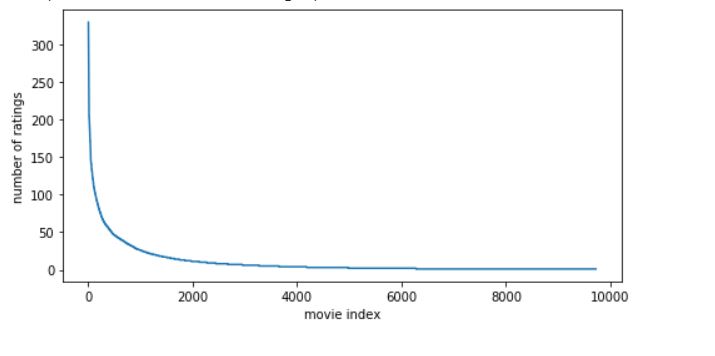
The sparsity index of the rating matrix is 0.0170.

## **Question 2.**

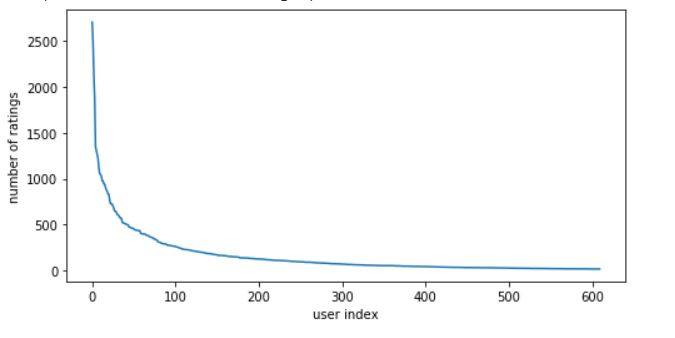


The distribution is left-skewed, and the overall ratings are high. People do it kindly.

## **Question 3.**



## **Question 4.**



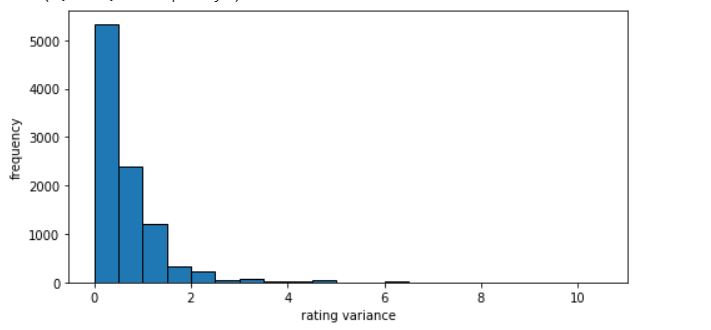
## **Question 5.**

Among all movies, only a small portion of movies received ratings from more than one user. Most of them received only one rating.

## 

## 

## **Question 6.**

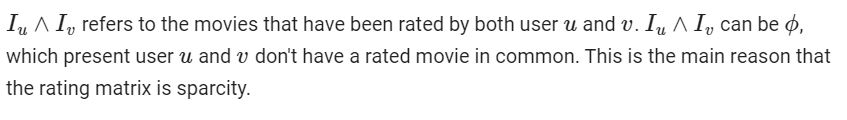


The distribution shows that the number of movies decreases with variance of rating values. This suggests that there are many movies that receive similar rating values.

## **Question 7.**



## **Question 8.**



## **Question 9.**

The mean-centering of the raw ratings helps to reduce the effect of bias by only considering the deviation from the mean. Generally, people may consistently rate items highly or lowly.

## 

## 

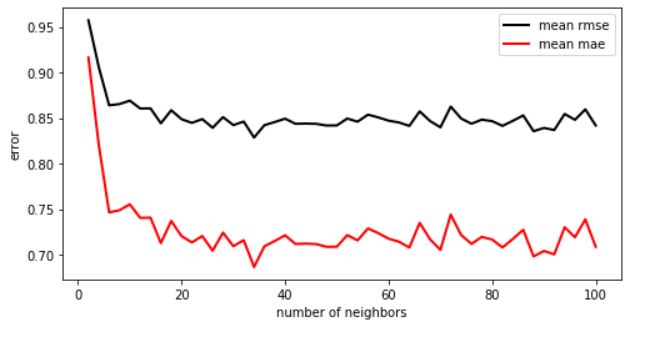
## 

## 

## 

## 

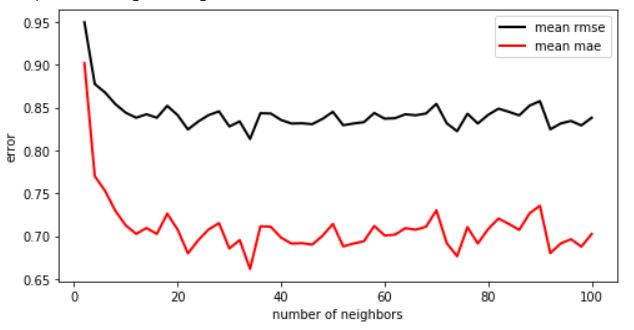
## **Question 10.**



## **Question 11.**

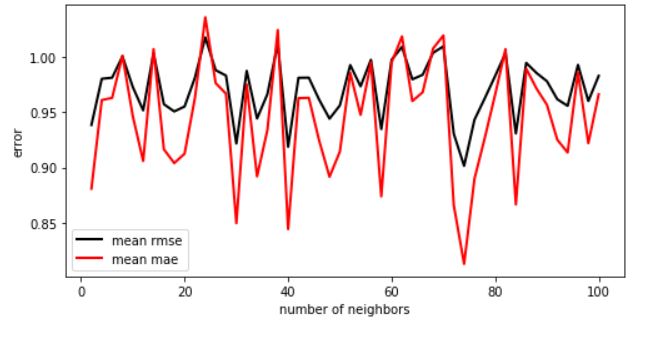
The minimum k is 20. The steady state MAE is 0.7157, and the steady state RMSE is 0.8460.

## **Question 12.**



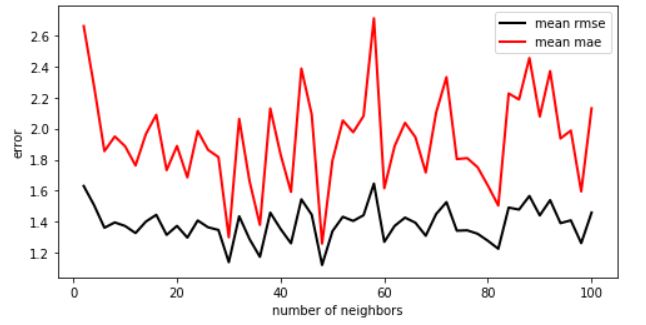
The mean RMSE is 0.83136, and the mean MAE is 0.69116.

## **Question 13.**



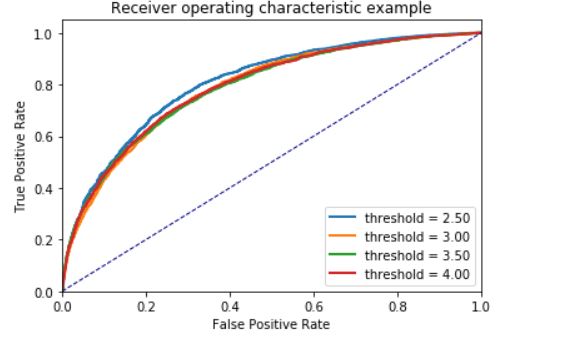
The mean RMSE is 0.90146, and the mean MAE is 0.812635.

## **Question 14.**



The mean RMSE is 1.387268, and the mean MAE is 1.93651.

## **Question 15.**

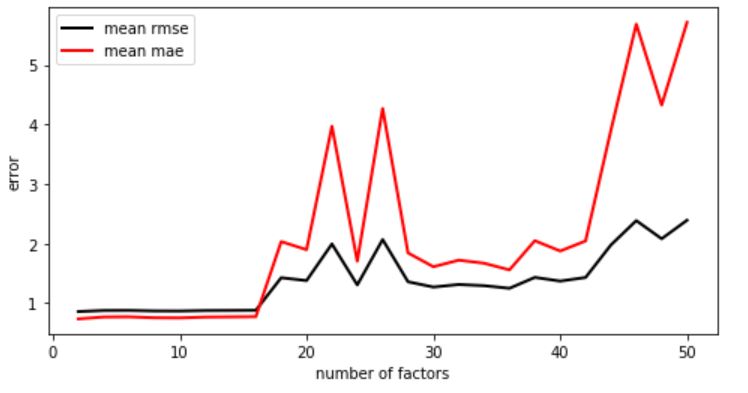


The AUC of 4 threshold are [0.8071771363933951, 0.787741919670092, 0.7871680867308565, 0.7899004020747654].

## **Question 16.**

The optimization problem is not convex because *U* and *V* are variables and the optimization problem involves the term u\_{i\*}v\_{j\*}^T. That is why we have to iteratively calculation. Each time we fixed *U* and solve *V*, then next time we fixed *V* and solve *U*.

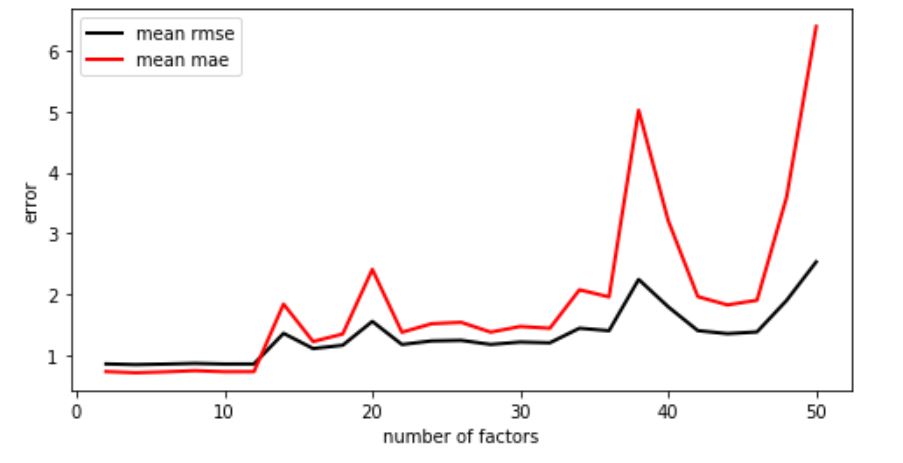
## **Question 17.**



## **Question 18.**

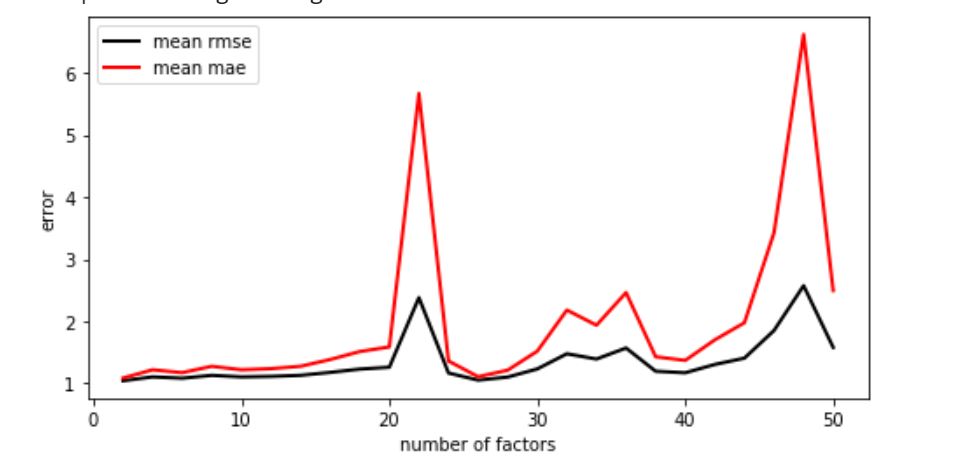
The optimal k should be 6. The minimum RMSE is 0.86516, and the minimum MAE is 0.7485. The optimal number of factor is not the same as the movie genres.

## **Question 19.**



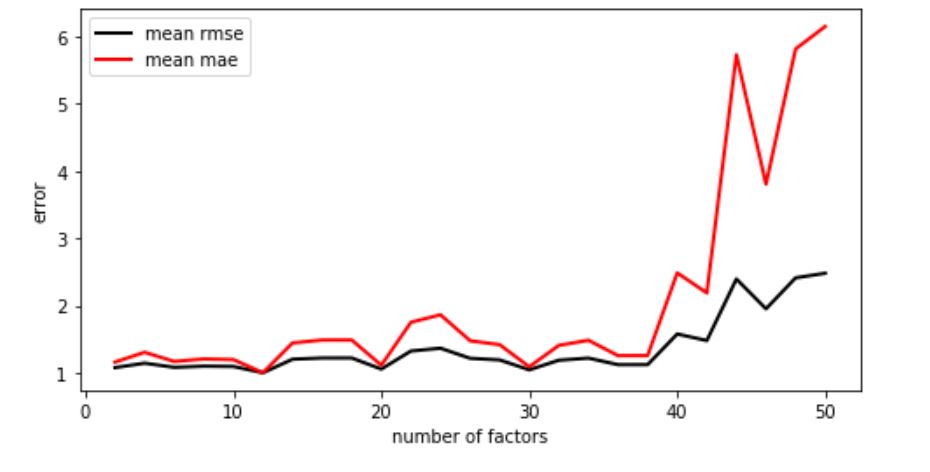
The minimum RMSE is 0.8406, and the minimum MAE is 0.70663.

## **Question 20.**



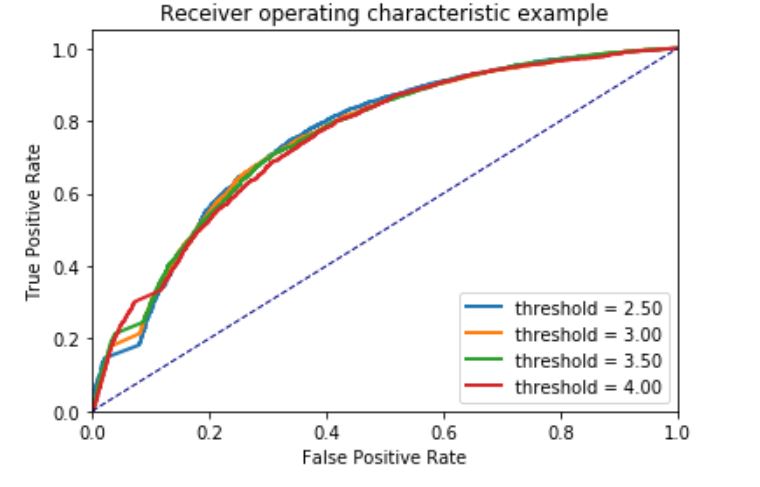
The minimum RMSE is 1.0416, and the minimum MAE is 1.0851.

## **Question 21.**



The minimum RMSE is 0.9999, and the minimum MAE is 0.9998.

## **Question 22.**



AUC for four thresholds are 0.7558597731845105, 0.7539469572053783, 0.75435606027009, and 0.7510821831577852.

## **Question 23.**

The top 10 movies for each k can be found that belong to a small collection of genres. The larger the k, the smaller difference between genres.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| K=5 | K=8 | K=10 | K=15 | K=20 |
| Comedy | Comedy|Horror|Sci-Fi | Drama|Musical | Documentary | Comedy|Drama|Romance |
| Drama|Mystery | Drama | Drama | Crime|Drama | Action|Animation|Horror |
| Drama | Action|Horror|Sci-Fi|Thriller | Drama | Action|Sci-Fi|Thriller | Drama |
| Adventure|Comedy|Fantasy | Comedy | Crime|Thriller | Comedy | Comedy |
| Drama | Documentary | Drama | Action|Adventure|Drama|Sci-Fi|Thriller | Action|Adventure|Thriller|IMAX |
| Drama|Film-Noir | Comedy | Animation|Fantasy|Mystery | Action|Drama|Thriller | Drama |
| Drama|Musical | Drama|War | Animation|Documentary | Comedy|Crime|Drama|Musical | Drama|Romance |
| Comedy|Crime | Comedy | Documentary|Drama | Action|Crime|Drama|Thriller | Comedy|Drama|Romance |
| Drama | Horror|Thriller | Comedy | Comedy | Drama|Horror|Thriller |
| Comedy|Mystery | Documentary|Musical | Comedy|Romance | Comedy | Comedy |

## 

## 

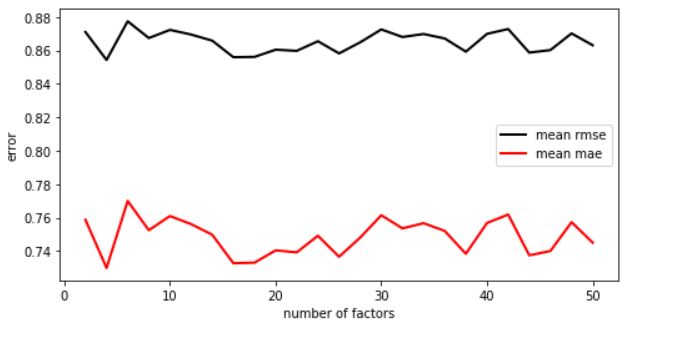
## 

## 

## 

## 

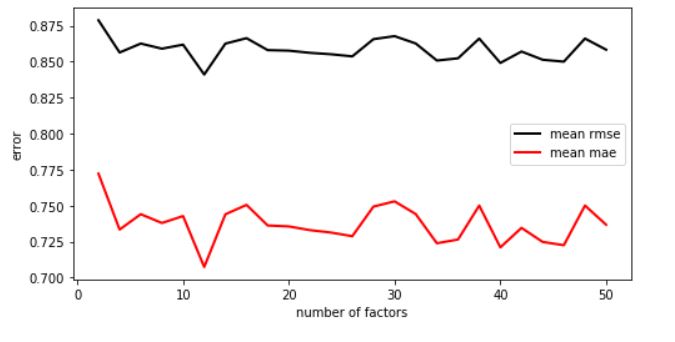
## **Question 24.**



## **Question 25.**

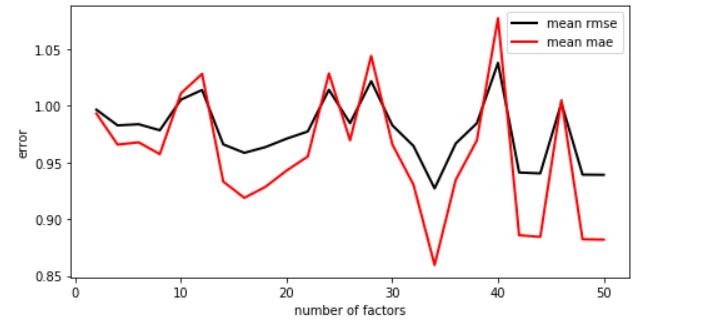
The best K is 1. The minimum RMSE is 0.8543064253417695, and the minimum MAE is 0.7298394683802323.

## **Question 26.**



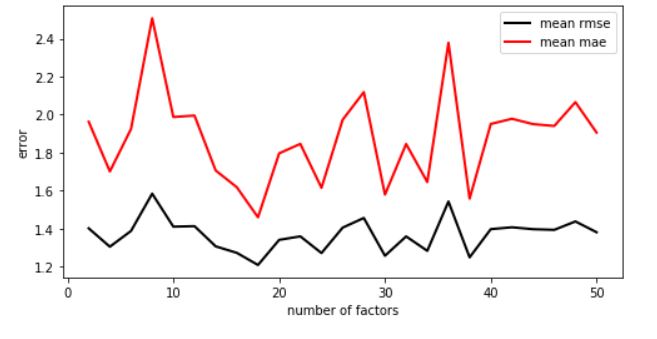
The minimum RMSE is 0.8410869443449591, and the minimum MAE is 0.7074272479475403.

## **Question 27.**



The minimum RMSE is0.9271348537695873, and the minimum MAE is 0.859579037074354.

## **Question 28.**

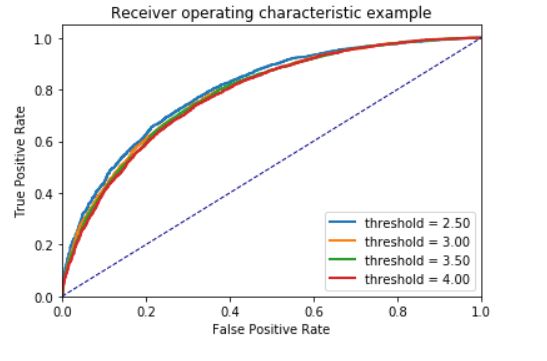


The minimum RMSE is 1.20783653022206, and the minimum MAE is 1.458869083738865.

## 

## 

## **Question 29.**



AUC are [0.7996821252348949, 0.7873291703983812, 0.7843557110139378, 0.7801965630715897], respectively.

## **Question 30.**

{'mean\_mse': **3.3405297086383636**, 'mean\_rmse': **3.5522342023797266,**

'mse\_hist': [3.352426243377296, 3.3405297086383636, 3.3355216505541017,

3.3366561018334555, 3.3472948468994743, 3.3462838014627314,

3.350853945675127, 3.3351887348678324, 3.349276236330121,

3.3211768011702056],

'rmse\_hist': [3.561840506612431, 3.5522342023797266,

3.549620006740572, 3.5461106184598044,

3.561606815312605, 3.5598752417702397,

3.5575782130465714, 3.54517016417337,

3.563271542594135, 3.5347993652047878]}

## **Question 31.**

{'mean\_mse': **3.350986388205029**, 'mean\_rmse': **3.5624200630861975**,

'mse\_hist': [3.3501967106283956, 3.350986388205029,

3.3374566079184564, 3.357701782042459, 3.340107398359884, 3.3619533597100206,

3.3418790953360467, 3.3621873497745742, 3.3360158561061786, 3.3430847741190974],

'rmse\_hist': [3.5612780825721013, 3.5624200630861975, 3.5511545163887406,

3.568866842743934, 3.5549658591417326, 3.572796844641089, 3.561648445070146,

3.575292585934002, 3.547347867752401, 3.5565346796258477]}

## **Question 32.**

{'mean\_mse': **0.765975315556907**, 'mean\_rmse': **0.9997054257294251**,

'mse\_hist': [0.8148933287336415, 0.765975315556907, 0.7345275653988816, 0.77086412175423,

0.7962002261416907, 0.7703281263139748, 0.7692652907567019, 0.7999632873402649,

0.7510400783620659, 0.7951742721967167],

'rmse\_hist': [1.0875860112436508, 0.9997054257294251, 0.9360661205569328,

1.0054746190137538, 1.034585900871471, 1.0495400925322207,

1.0096012492307467, 1.0520261398461157, 1.0059606299242305, 1.1163527768990111]}

## **Question 33.**

{'mean\_mse': **1.2073732916984454**, 'mean\_rmse': **1.6150685925287664**,

'mse\_hist': [1.1172064608612293, 1.2073732916984454, 0.9719613637791761,

1.1747884069718746, 1.0882521530959879, 1.1299496440678933,

1.178022400764197, 0.9892896688228854, 1.3384616510672305,

1.1853921492175394],

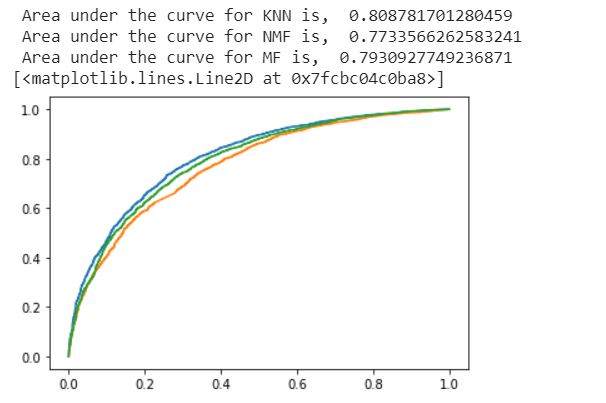
'rmse\_hist': [1.4329787452156835, 1.6150685925287664,

1.2261890522136918, 1.4868804342378517, 1.3610538943435169,

1.3686406739755181, 1.4803595813969181, 1.3197322906245583,

1.6333547300003497, 1.4453067056534157]}

## **Question 34.**



The KNN performs the best.

## **Question 35.**

Precision means the percentage of your results which are relevant. On the other hand, recall refers to the percentage of total relevant results correctly classified by your algorithm.

## **Question 36.**

## **Question 37.**

## **Question 38.**

## **Question 39.**

## 