Project3

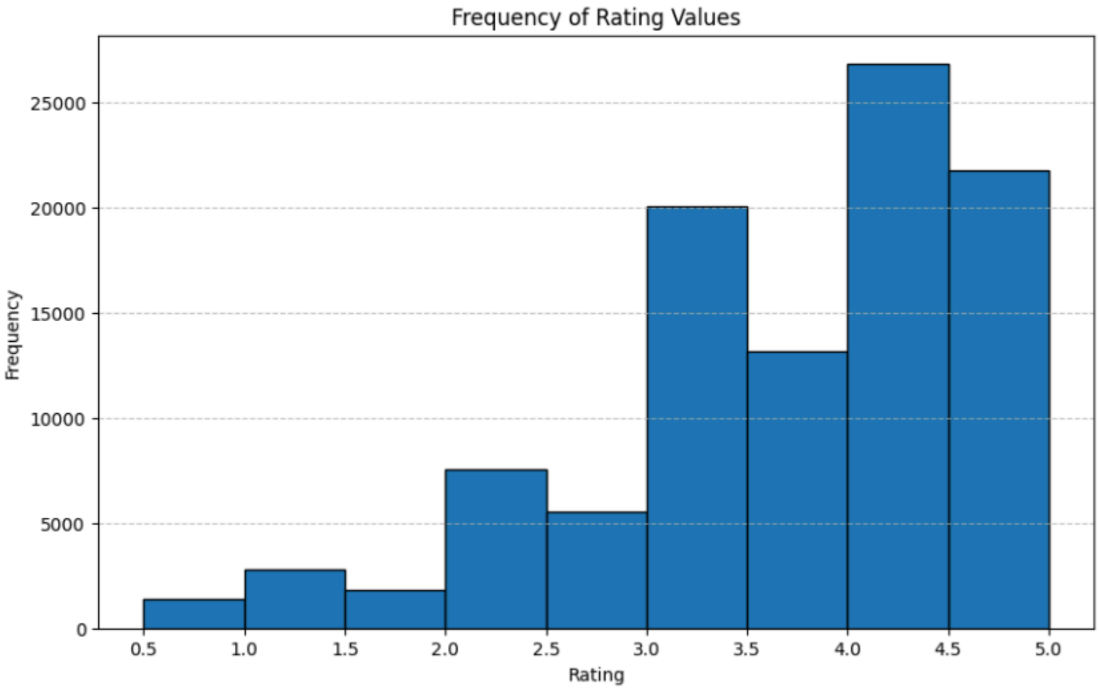
Hexi Meng (406200552),Zhanhong Liu(206152835)

QUESTION 1:

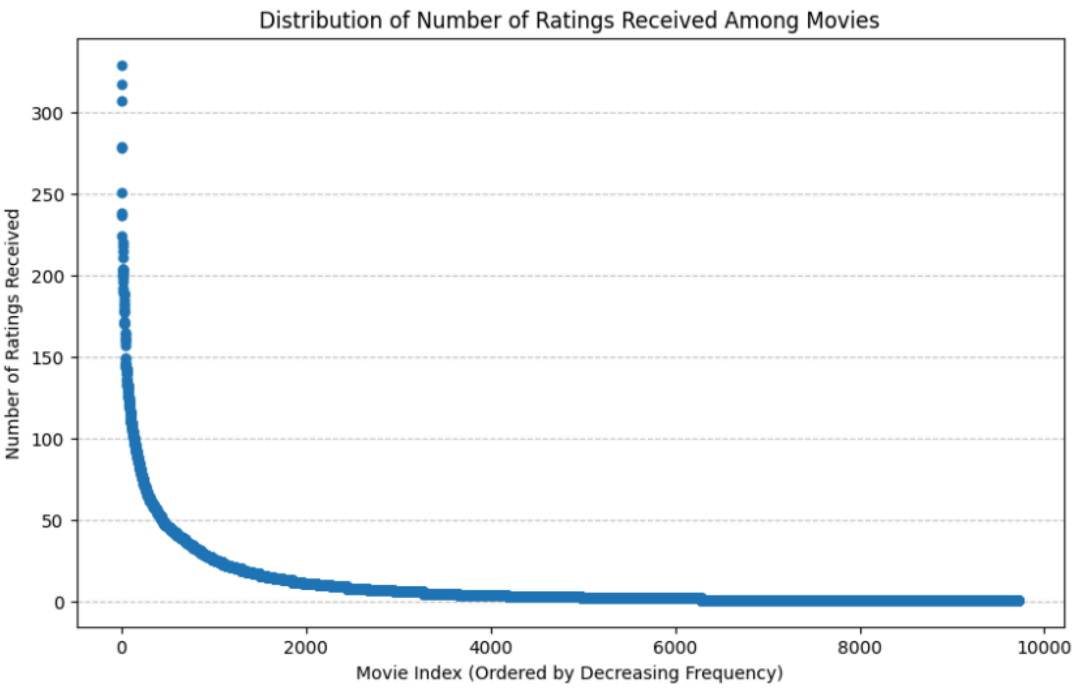
A

Sparsity = 0.016999683055613623

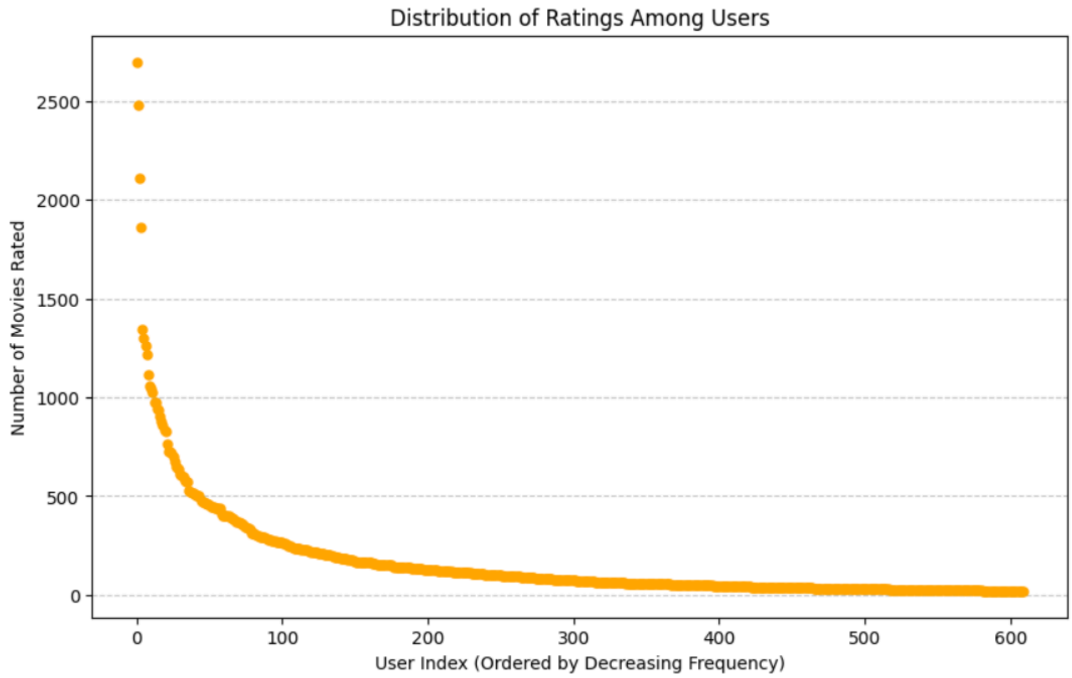
B



C



D

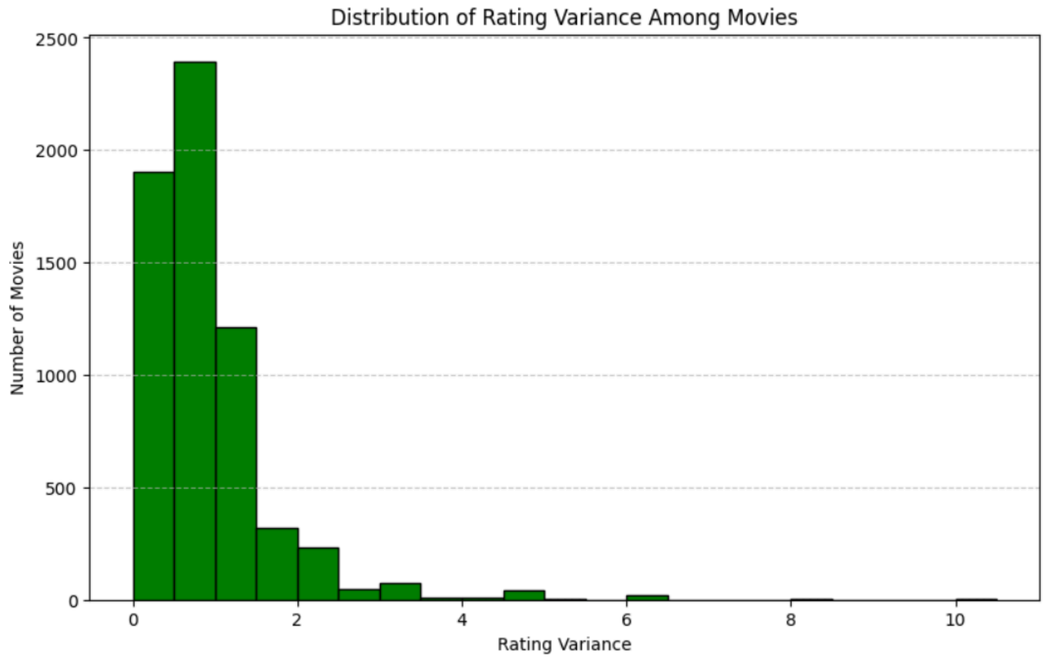


E

The key features from the distributions indicate:

1. **Long-tail distribution:** There are a few highly popular movies, suggesting the need for recommendations to balance between popular and niche movies for diversity.
2. **User engagement variability:** With users varying widely in the number of movies rated, recommendation systems must adapt to both active and less active users.
3. **Data sparsity:** The sparse nature of the dataset highlights the need for advanced techniques to accurately predict user preferences.

F



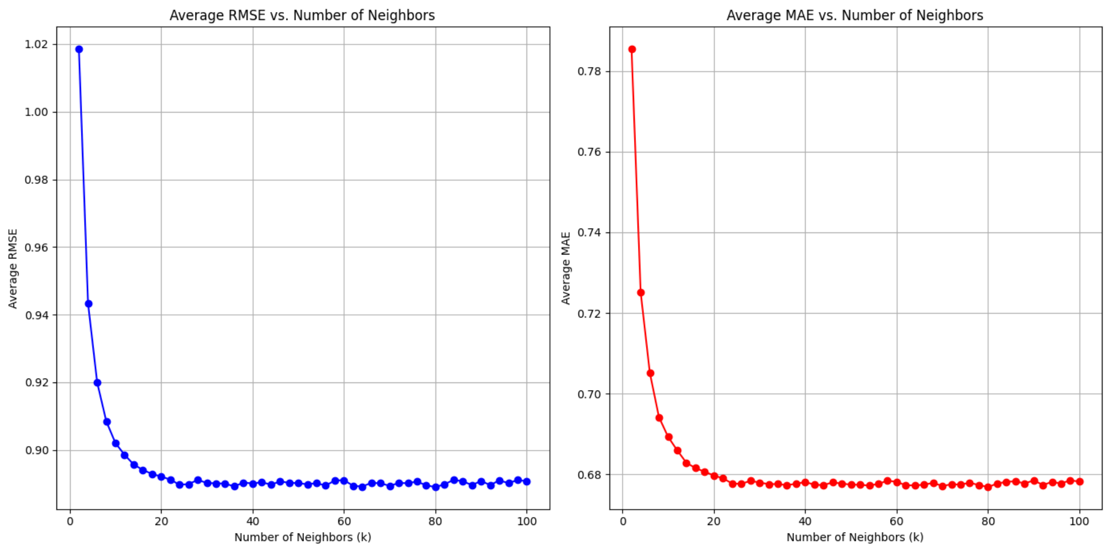
Question 2
A

$$\mu_u = \frac{1}{|I_u|} \sum_{k \in I_u} r_{uk}$$

B
The term $I_u \cap I_v$ represents the intersection of the sets of item indices for which users u and v have both provided ratings. Given that the rating matrix R is sparse, it is entirely possible that $I_u \cap I_v = \emptyset$.

Question 3
Mean-centering the raw ratings $(r_{vj} - \mu_v)$ The prediction function helps to adjust for individual user biases in their rating scales, ensuring that the prediction reflects genuine preferences rather than skewed high or low ratings. This approach allows the model to accurately capture and compare the relative likes and dislikes of users, improving the quality of recommendations.

Question 4

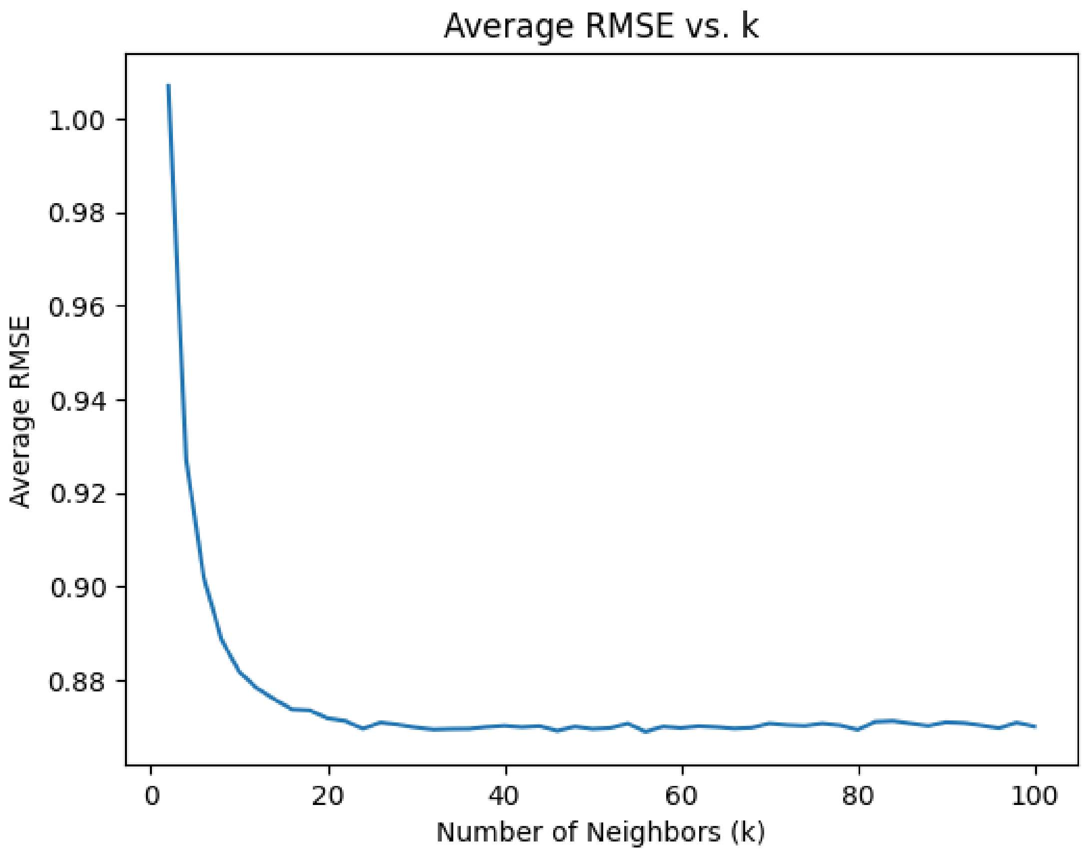


Question 5

$$minimum\ K = 22$$
$$average\ RMSE = 0.8912371255664379$$
$$average\ MAE = 0.6790317342207222$$

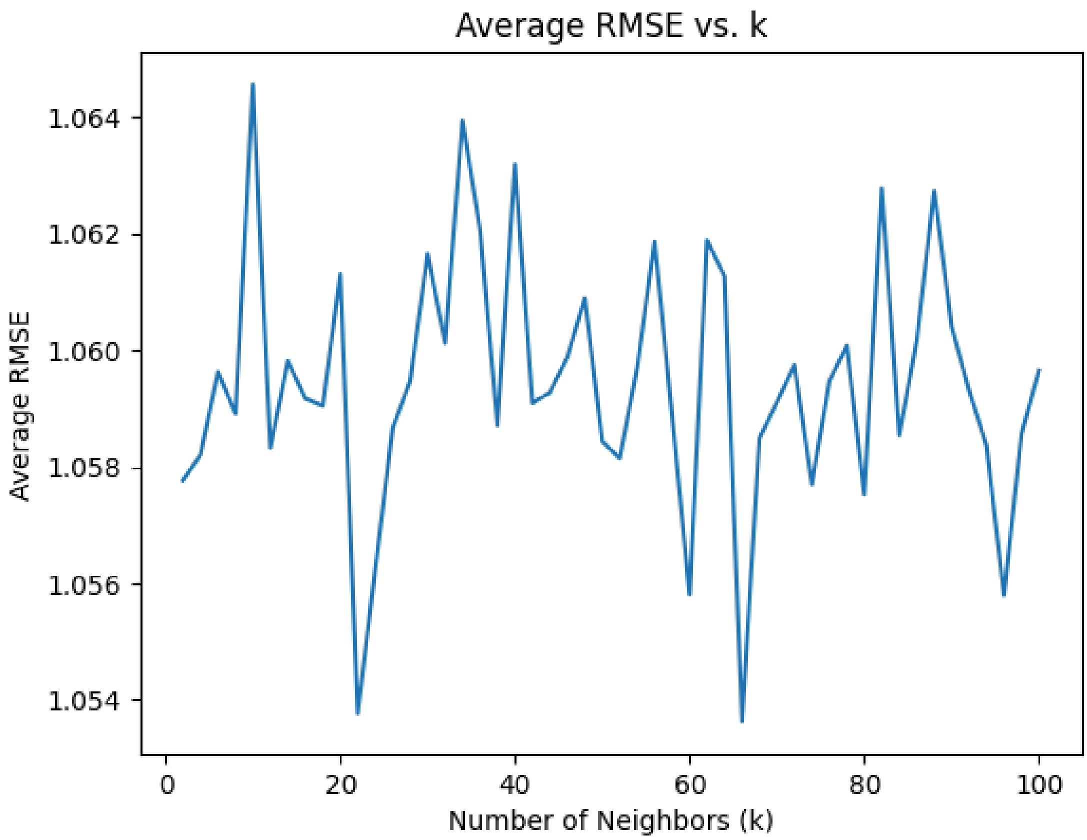
Question 6

Popular



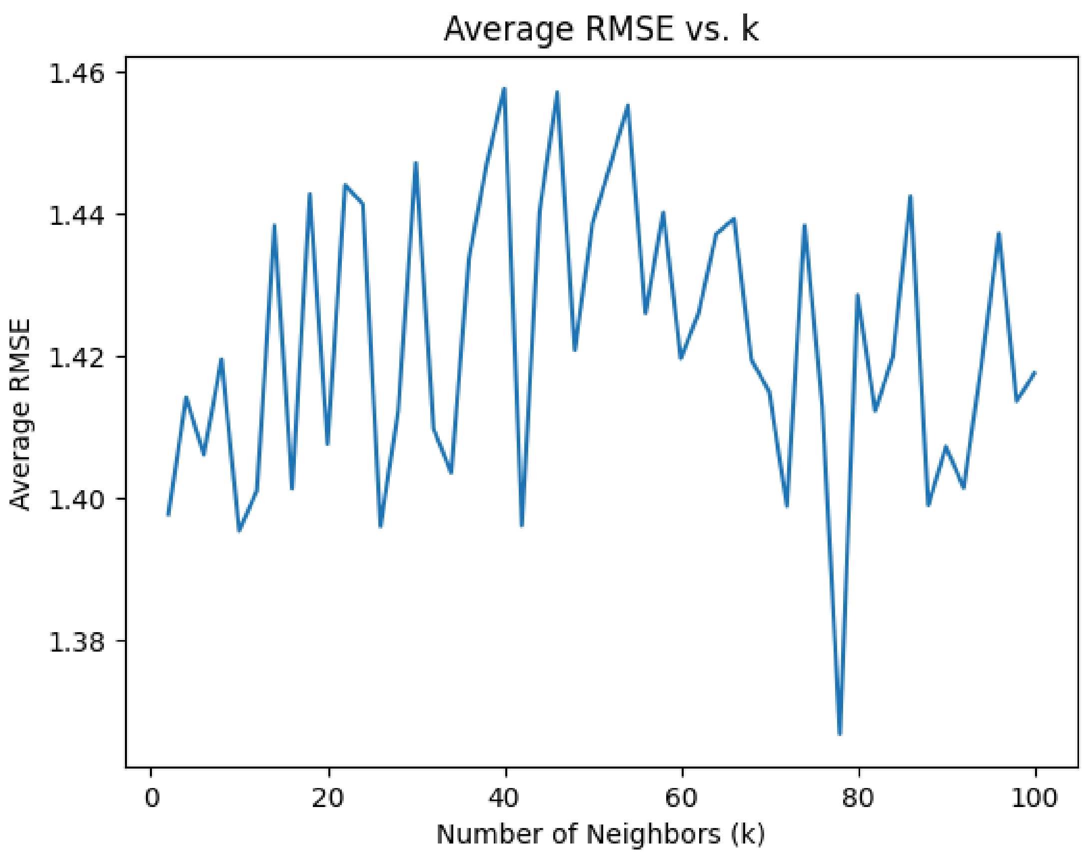
$$minimum\ average\ RMSE = 0.8689124913188335$$

Unpopular



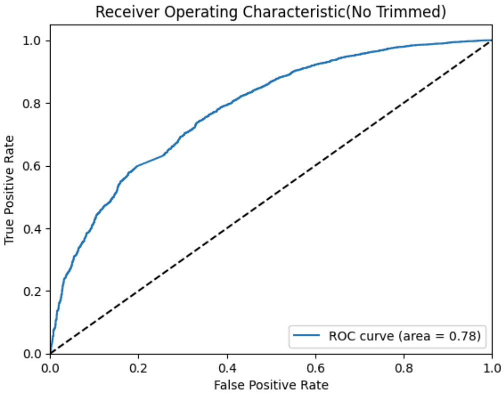
$$minimum\ average\ RMSE\ =\ 1.053621608103858$$

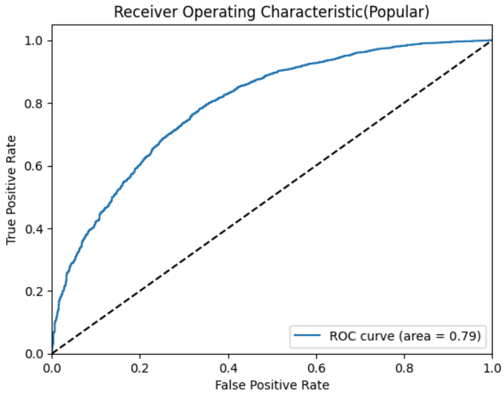
High-Variance

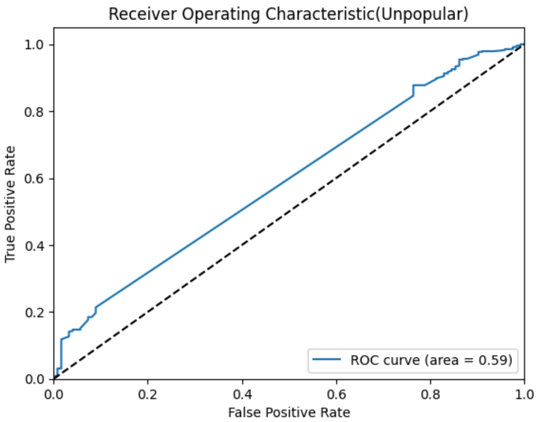


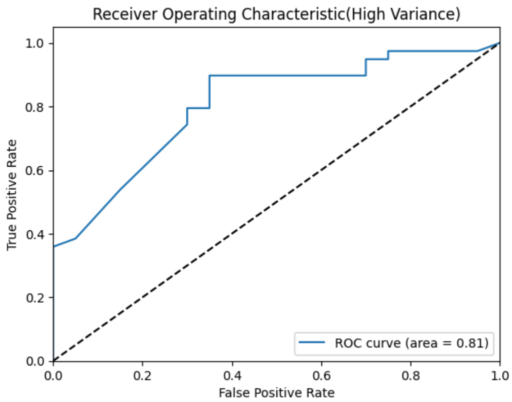
$$minimum\ average\ RMSE\ =\ 1.3668581671348572$$

Threshold = 2.5

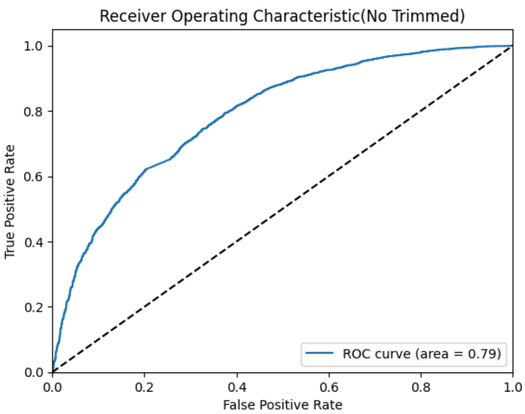


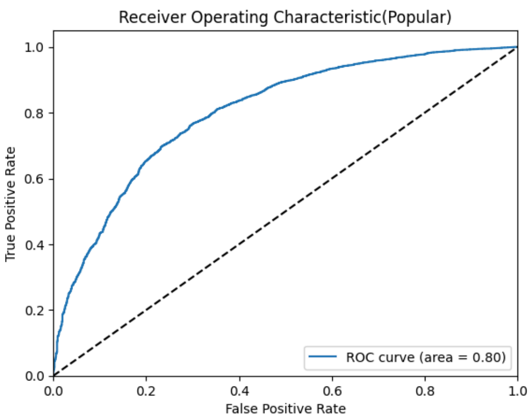


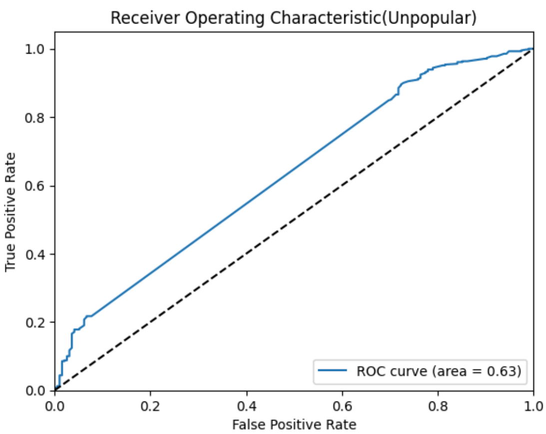


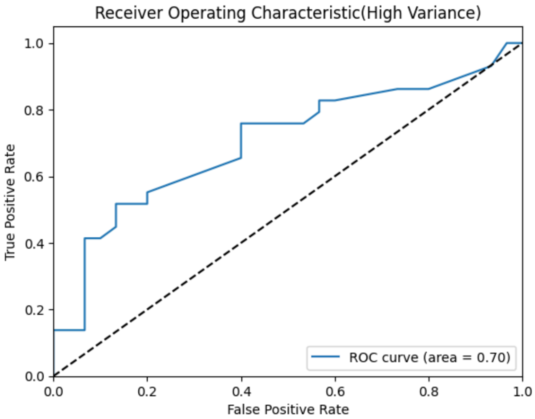


Threshold = 3

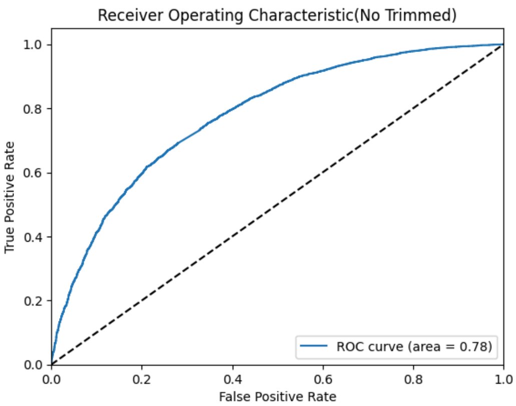


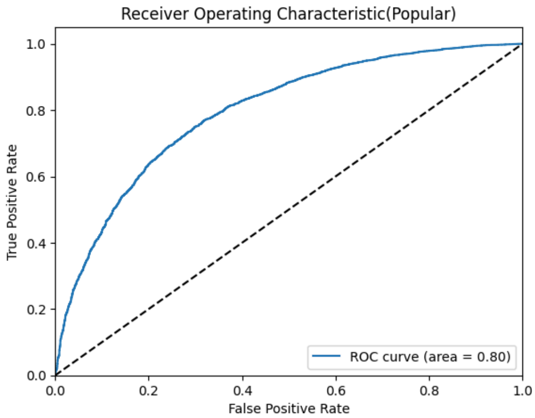


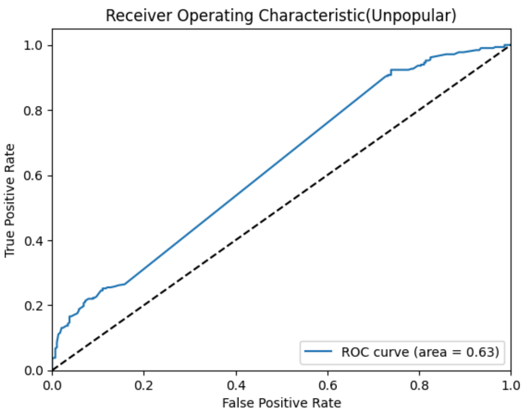


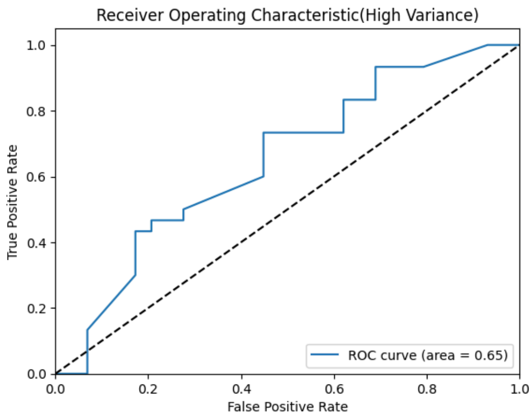


Threshold = 3.5

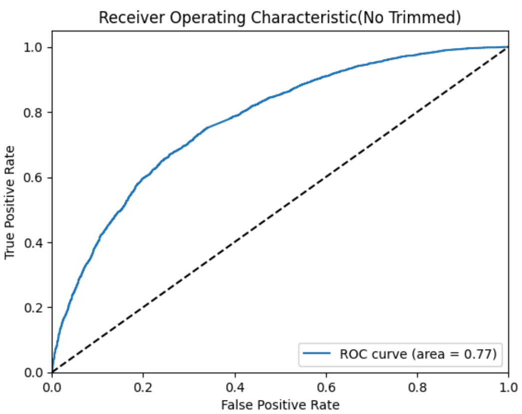


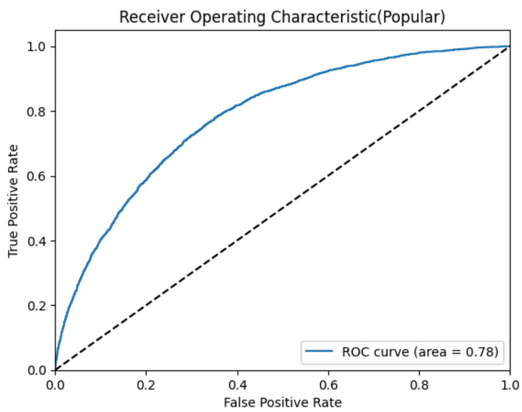


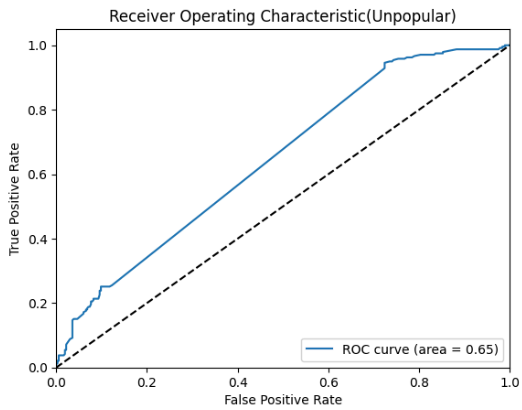


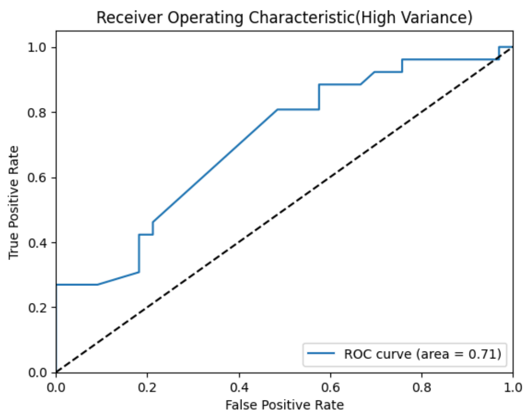


Threshold = 4









Question 7

The NMF optimization problem is not jointly convex in U and V. When either *U* or *V* is held fixed, the problem becomes convex in the other variable.

When *U* is fixed, the objective function becomes:

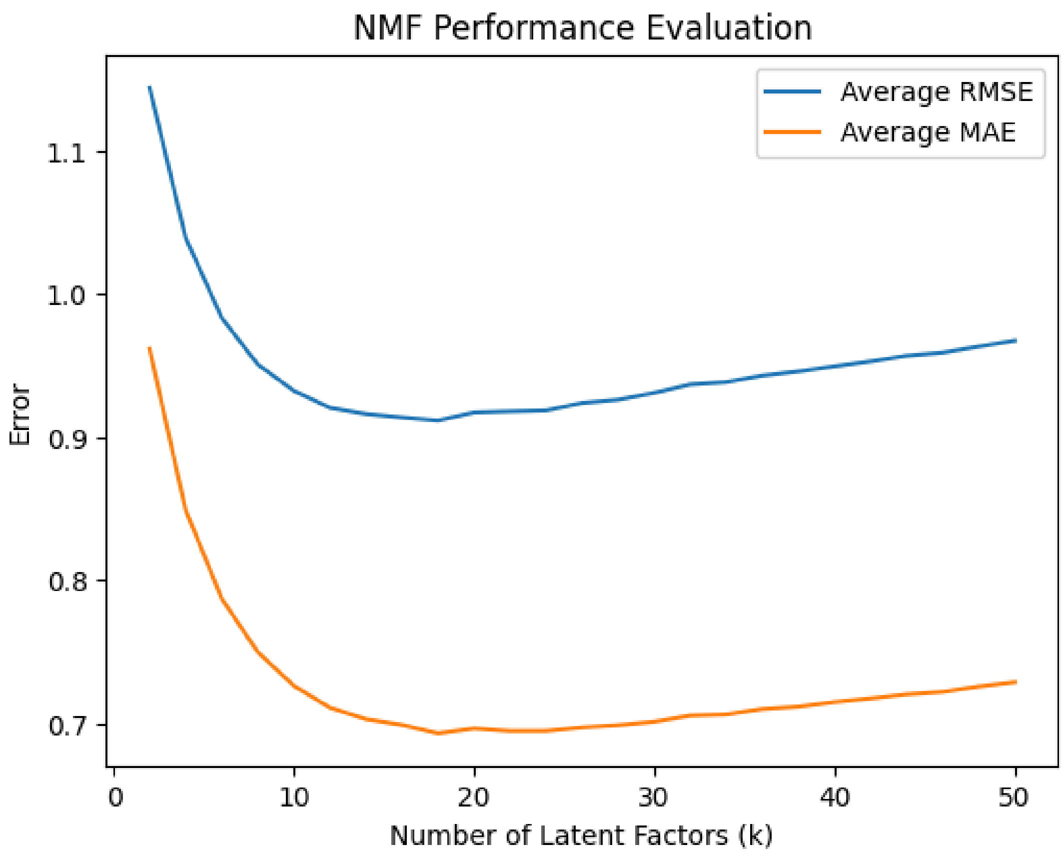
$$\sum_{i=1}^{m}\sum_{j=1}^{n} W_{ij}\left(r_{ij} - \left(UV^{T}\right)_{ij}\right)^{2} + \lambda\|V\|_{F}^{2}$$

$$Subject\ to\ V \geq 0$$

This can be viewed as a least-squares problem where you are trying to find *V* that minimizes the squared Frobenius norm of the difference between the observed ratings *R* and the product UV^{T}, with an added regularization term $\lambda\|V\|_{F}^{2}$ to prevent overfitting.

Question 8

A

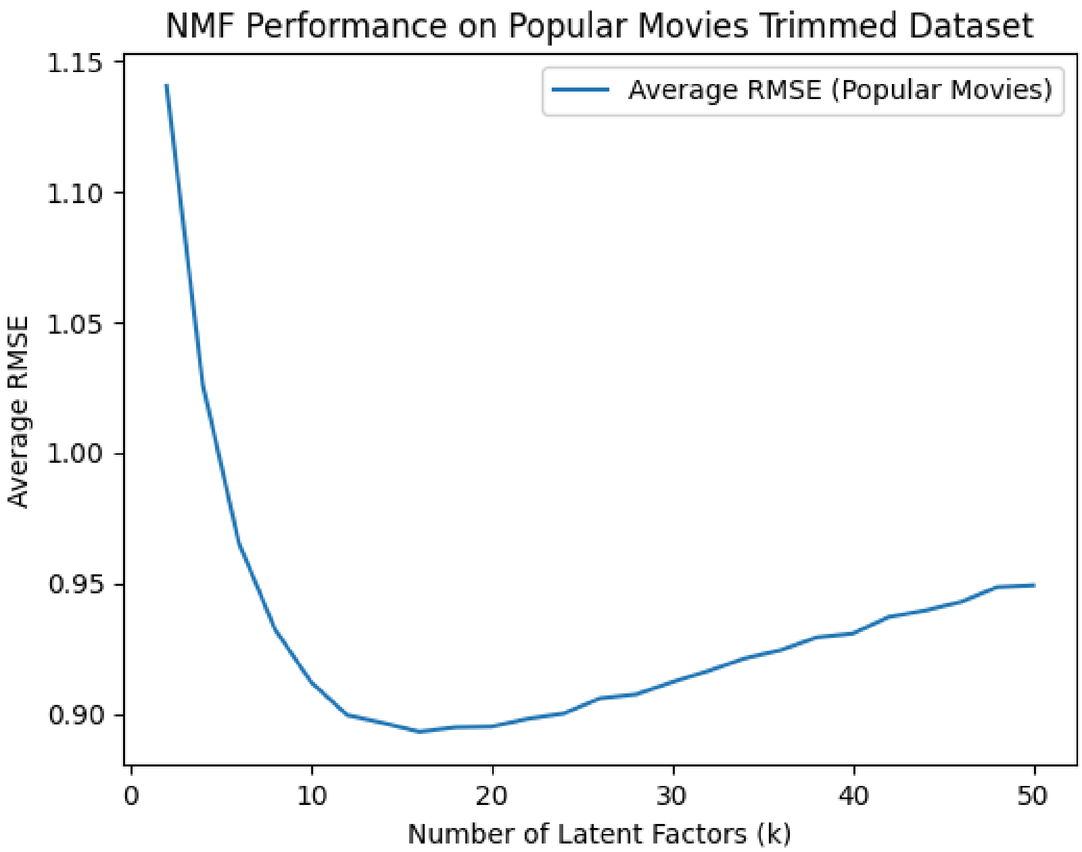


B

$$Optimal\ number\ of\ latent\ factors\ =\ 18$$
$$minimum\ average\ RMSE\ =\ 0.9116506399103693$$
$$minimum\ average\ MAE\ =\ 0.6932552267288592$$
$$the\ number\ of\ movie\ genres\ =\ 19$$

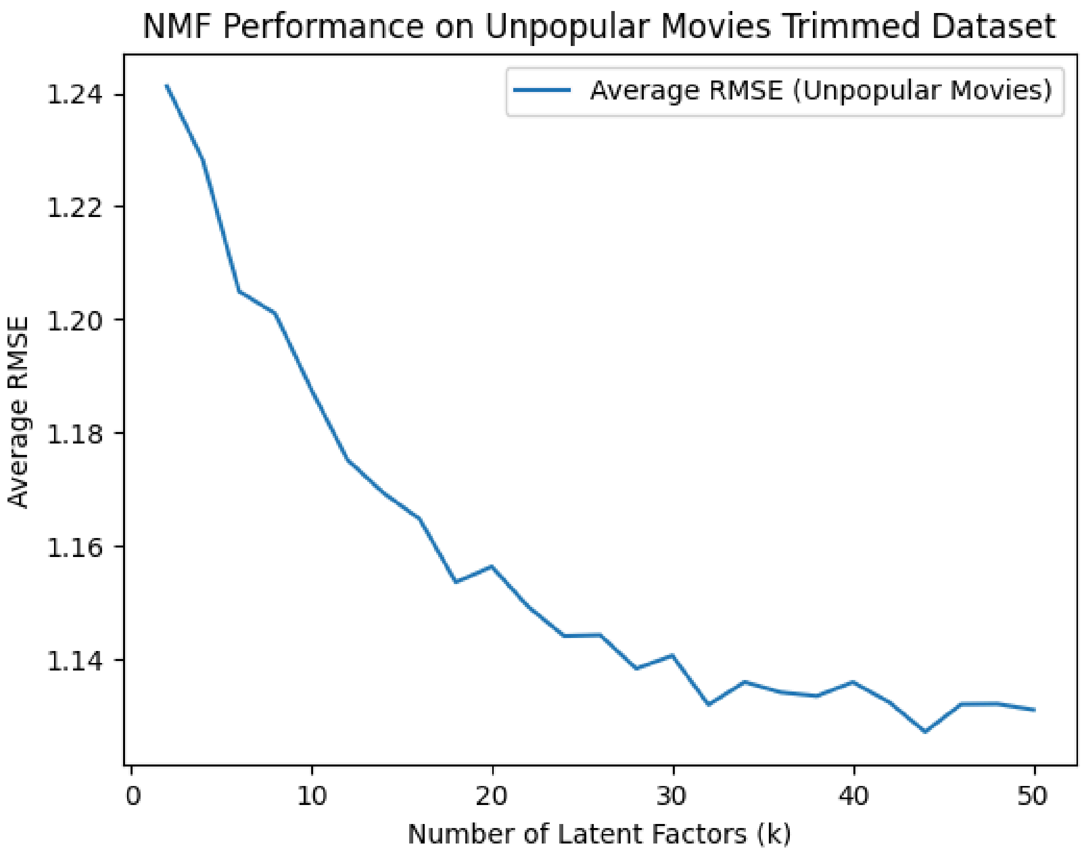
C

Popular



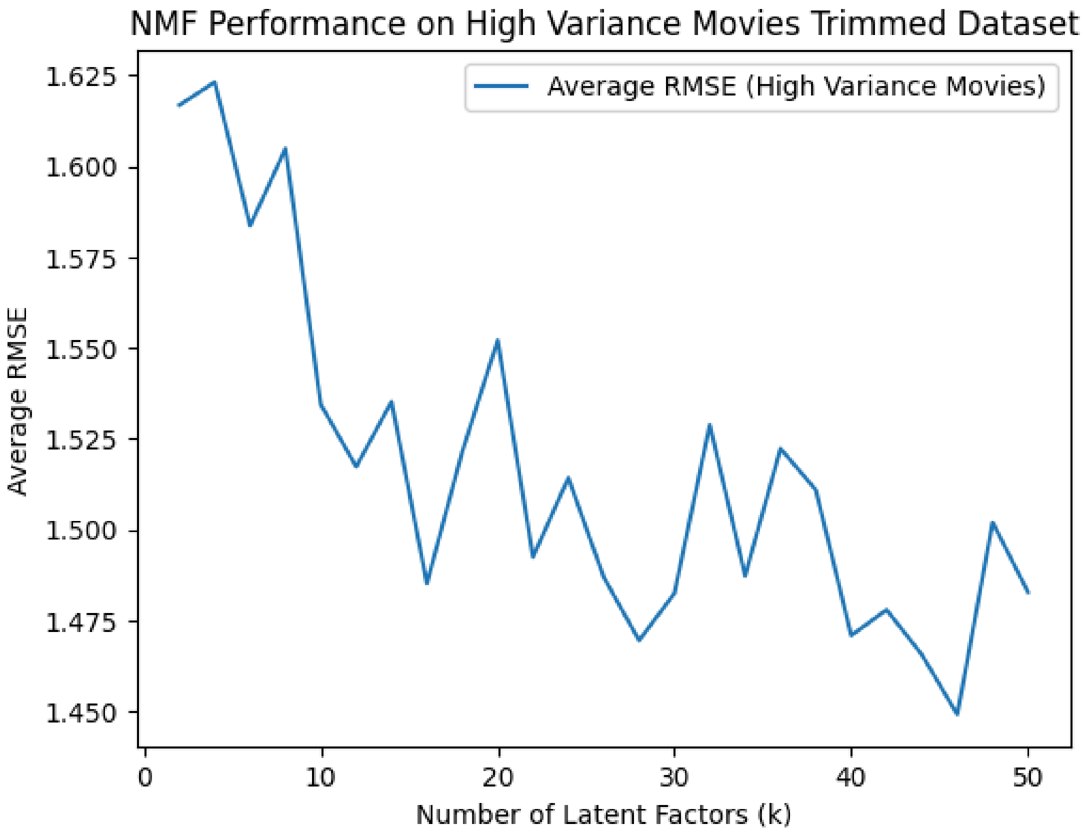
$$minimum\ average\ RMSE\ =\ 0.8931547399384527$$

Unpopular



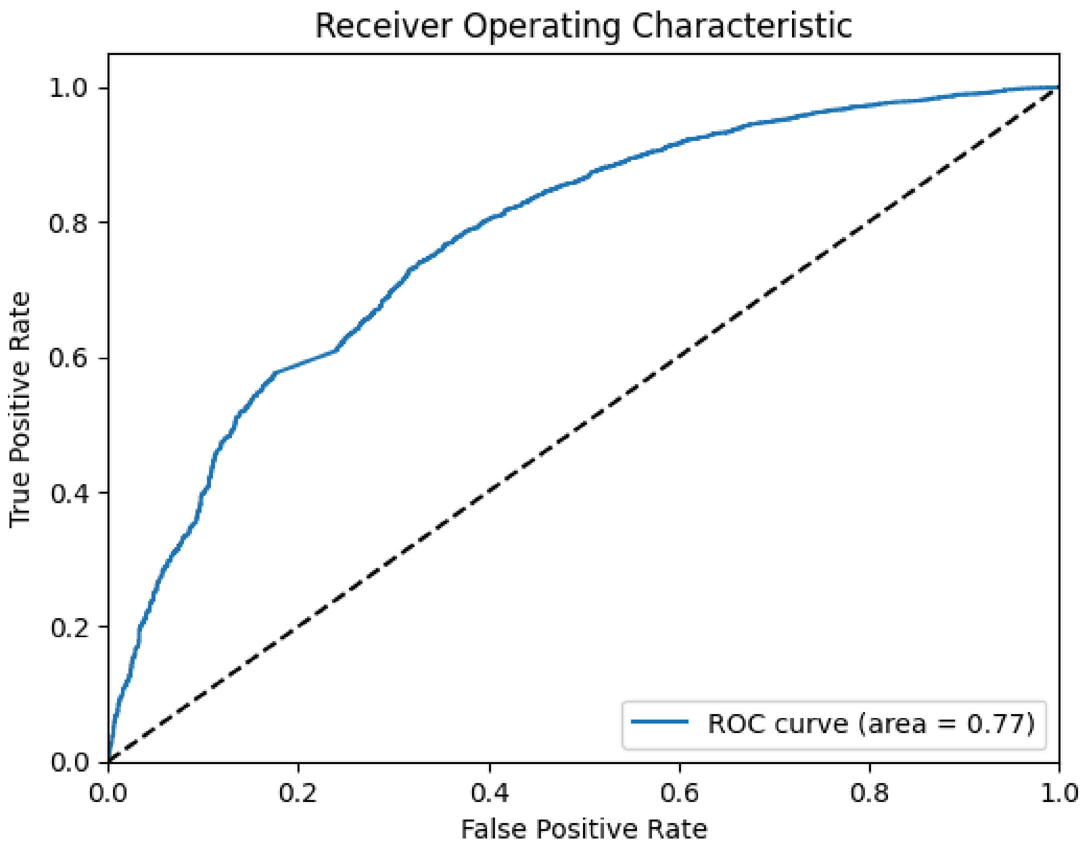
$$minimum\ average\ RMSE \ = \ 1.1270824832464257$$

High-Variance

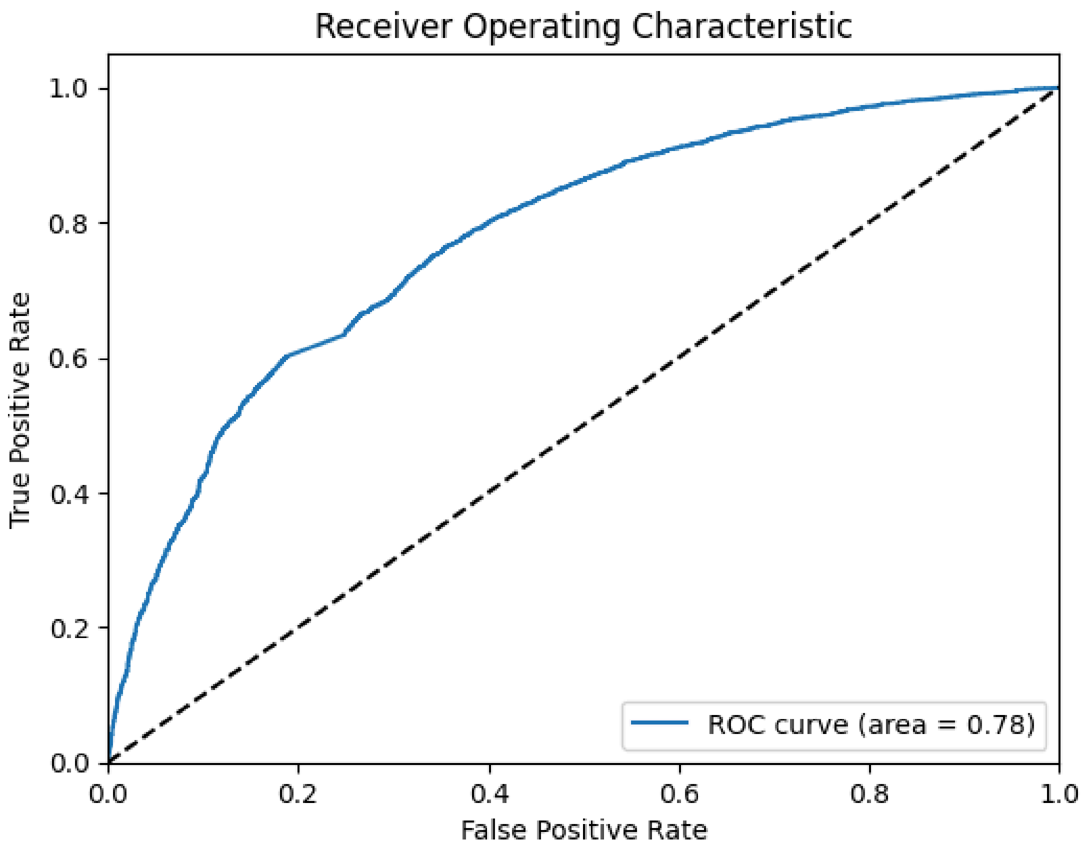


$$minimum\ average\ RMSE \ = \ 1.4491719193285797$$

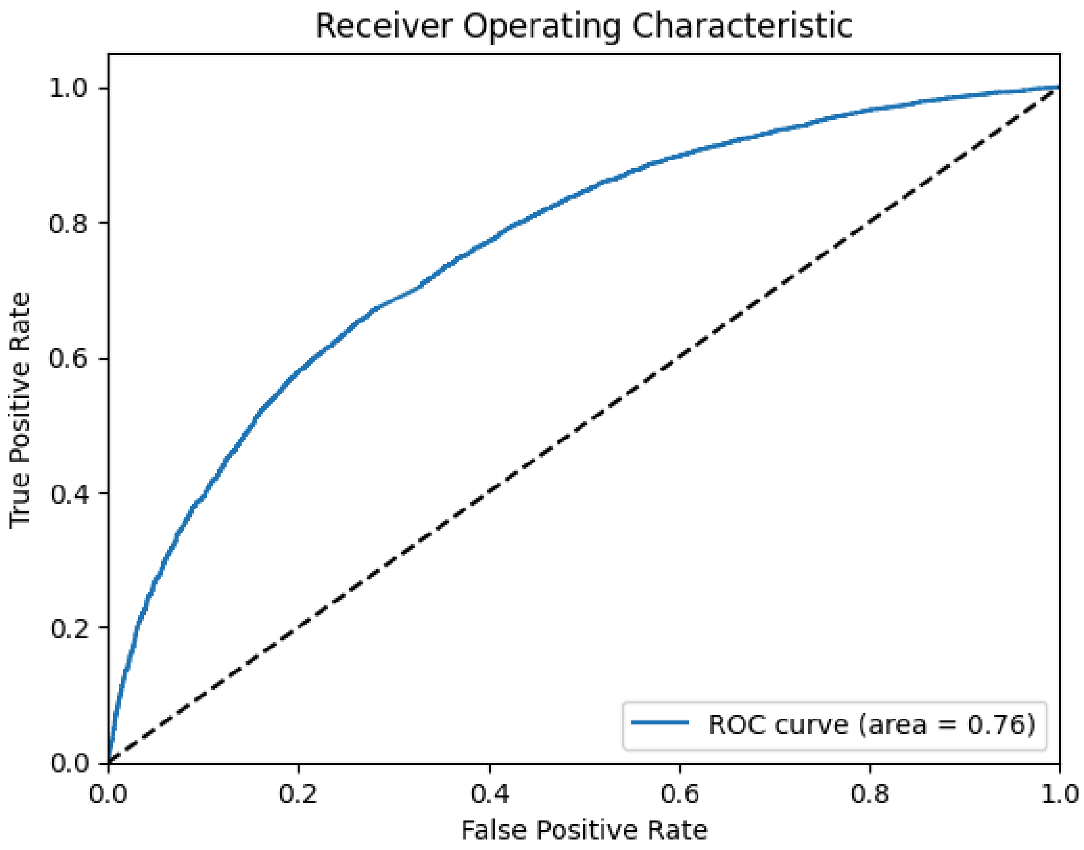
Threshold = 2.5



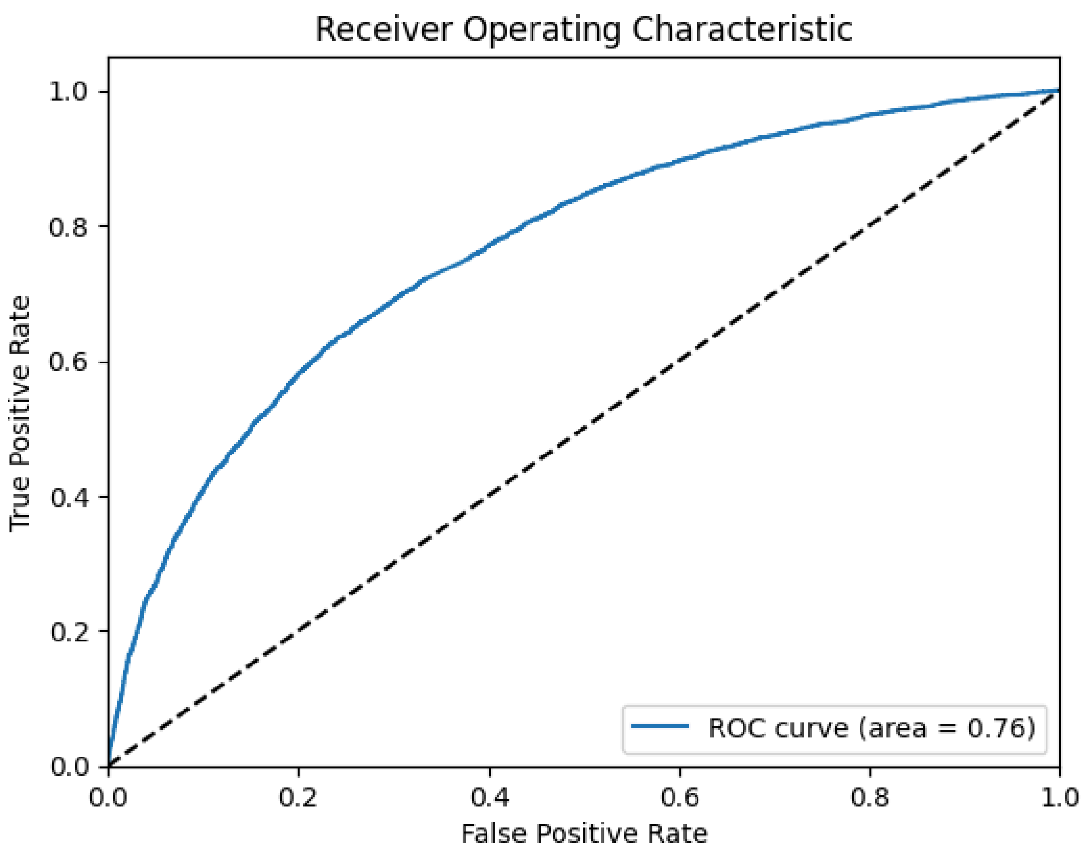
Threshold = 3.0



Threshold = 3.5



Threshold = 4.0



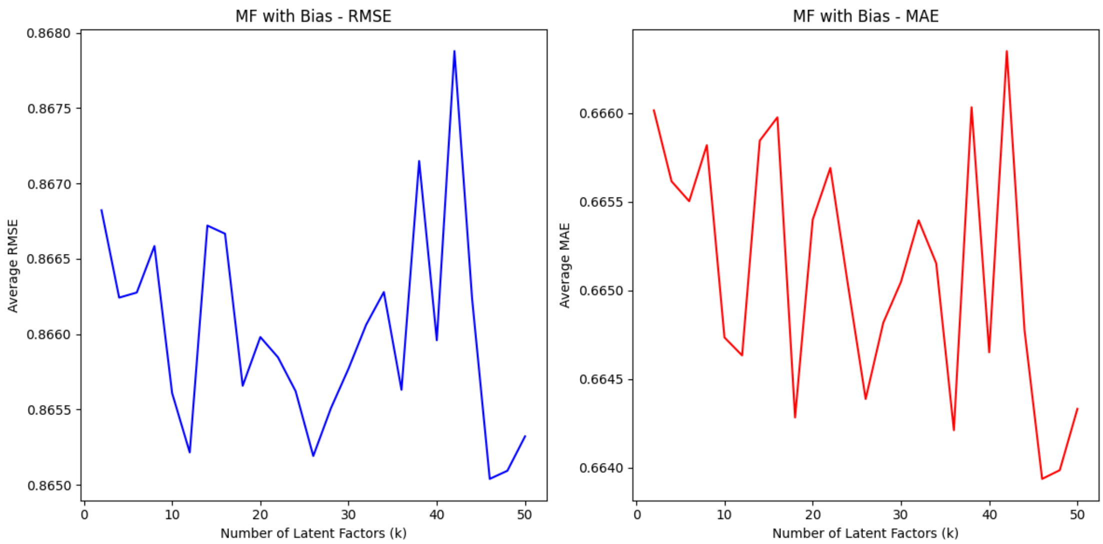
QUESTION 9

```
Top 10 Values for Latent Factor 0:
It's Such a Beautiful Day (2012): Animation|Comedy|Drama|Fantasy|Sci-Fi
Polar Express, The (2004): Adventure|Animation|Children|Fantasy|IMAX
Barbarella (1968): Adventure|Comedy|Sci-Fi
Dragon Ball Z: The History of Trunks (Doragon bôru Z: Zetsubô e no hankô!! Nokosareta chô senshi - Gohan to Torankusu) (1993): Action|Adventure|Animation
Peter Pan (2003): Action|Adventure|Children|Fantasy
Neon Genesis Evangelion: Death & Rebirth (Shin seiki Evangelion Gekijô-ban: Shito shinsei) (1997): Action|Animation|Mystery|Sci-Fi
Dead Ringers (1988): Drama|Horror|Thriller
Never Let Me Go (2010): Drama|Romance|Sci-Fi
Muse, The (1999): Comedy
Gothika (2003): Horror|Thriller
Top 10 Values for Latent Factor 1:
Troll 2 (1990): Fantasy|Horror
Master of the Flying Guillotine (Du bi quan wang da po xue di zi) (1975): Action
Piranha (1978): Horror|Sci-Fi
Crash (1996): Drama|Thriller
Dragon Ball Z the Movie: The Tree of Might (Doragon bôru Z 3: Chikyû marugoto chô kessen) (1990): Action|Adventure|Animation|Sci-Fi
Inland Empire (2006): Drama|Mystery|Thriller
Peeping Tom (1960): Drama|Horror|Thriller
Kwaidan (Kaidan) (1964): Horror
Hangar 18 (1980): Action|Sci-Fi|Thriller
Clonus Horror, The (1979): Horror|Sci-Fi
Top 10 Values for Latent Factor 2:
Joy Ride (2001): Adventure|Thriller
Gulliver's Travels (1939): Adventure|Animation|Children
UHF (1989): Comedy
Spitfire Grill, The (1996): Drama
Chocolat (1988): Drama
Rules of Attraction, The (2002): Comedy|Drama|Romance|Thriller
My Boss's Daughter (2003): Comedy|Romance
Rugrats in Paris: The Movie (2000): Animation|Children|Comedy
Armour of God II: Operation Condor (Operation Condor) (Fei ying gai wak) (1991): Action|Adventure|Comedy
Peggy Sue Got Married (1986): Comedy|Drama
```

The top 10 movies associated with a particular latent factor in an NMF model often belong to specific or a small collection of genres. This indicates a connection between the latent factors and movie genres, where each latent factor may represent underlying themes or genre characteristics common among certain movies.

Question 10

A

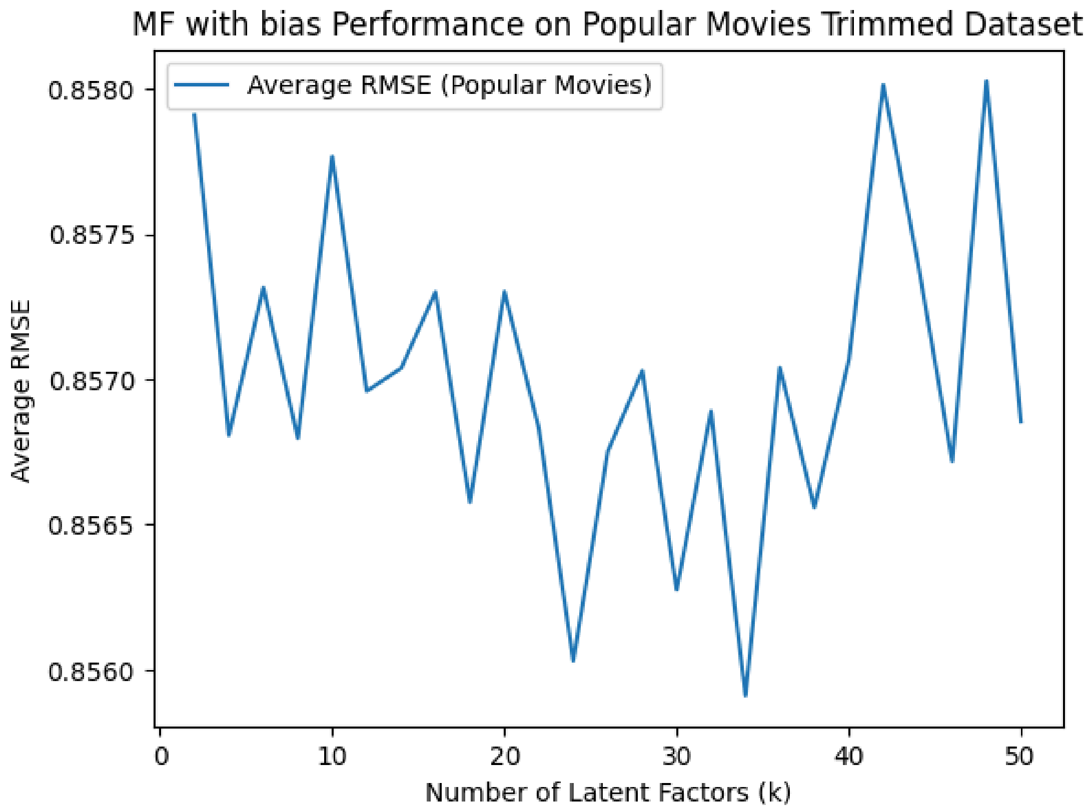


B

$$Optimal\ number\ of\ latent\ factors\ = 46$$
$$minimum\ average\ RMSE\ =\ 0.8650394896157966$$
$$minimum\ average\ MAE\ =\ 0.6639358146586974$$
$$the\ number\ of\ movie\ genres\ = 19$$

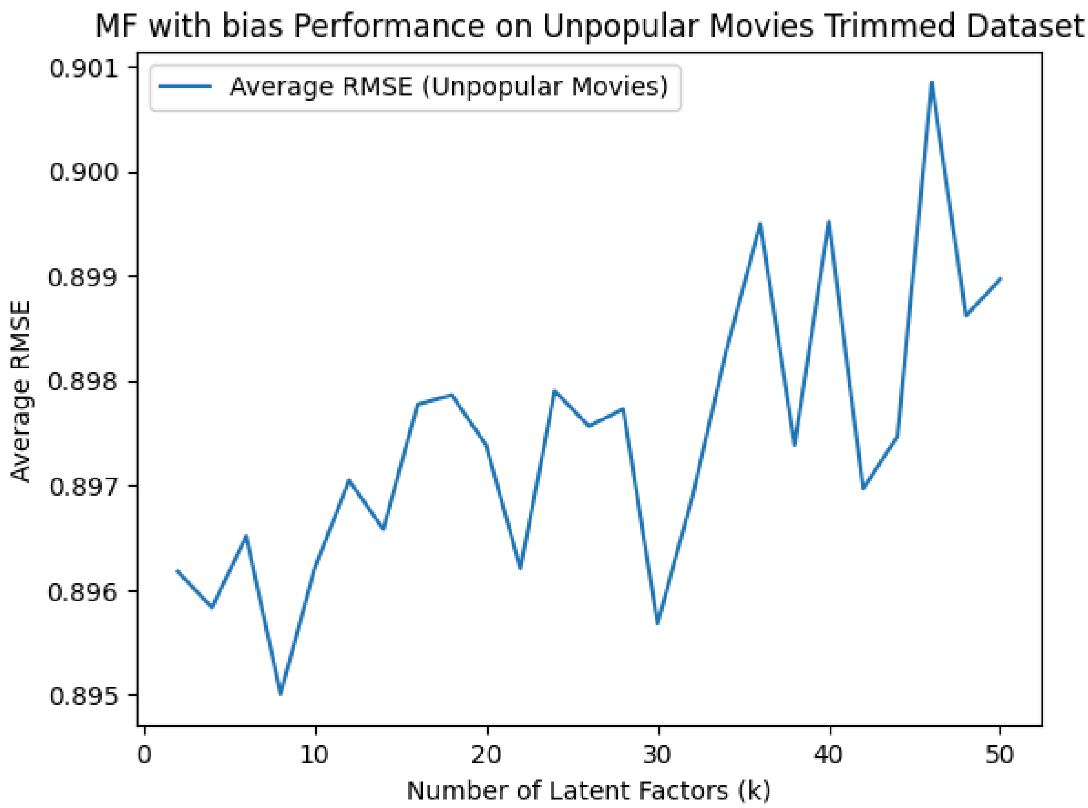
C

Popular



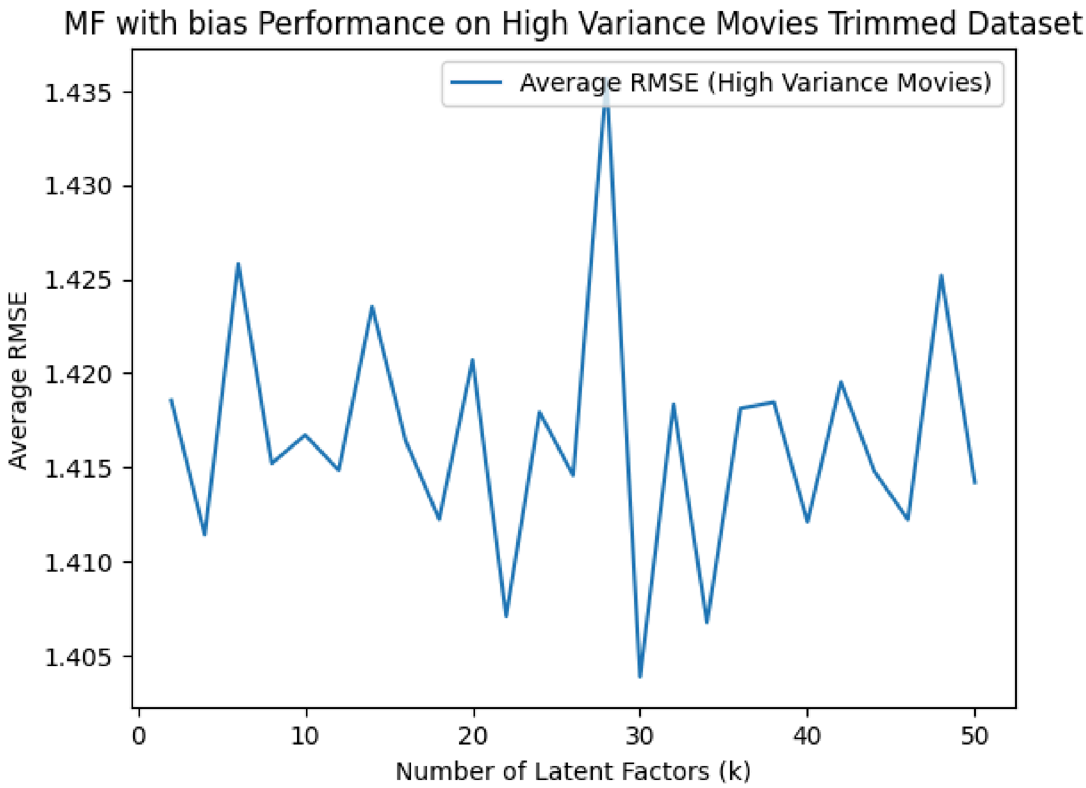
$$minimum\ average\ RMSE\ =\ 0.8559110478080288$$

Unpopular



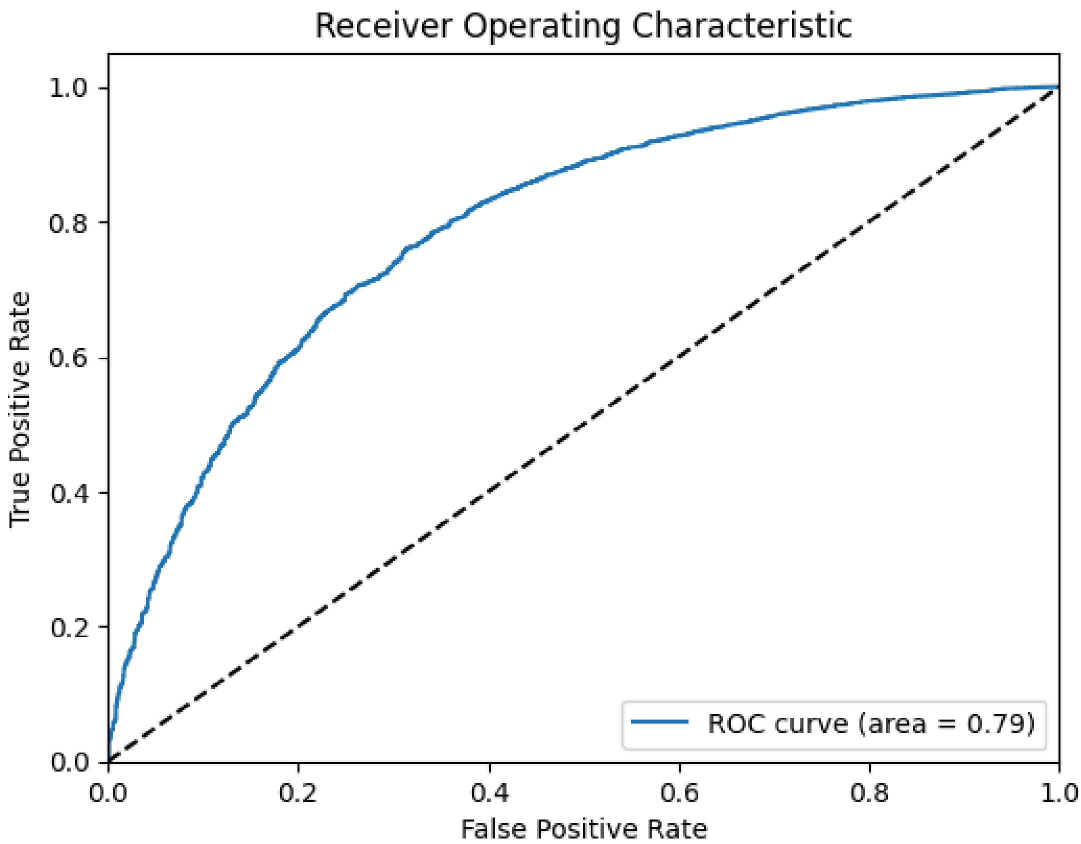
$$minimum\ average\ RMSE\ =\ 0.8950073913298912$$

High-Variance

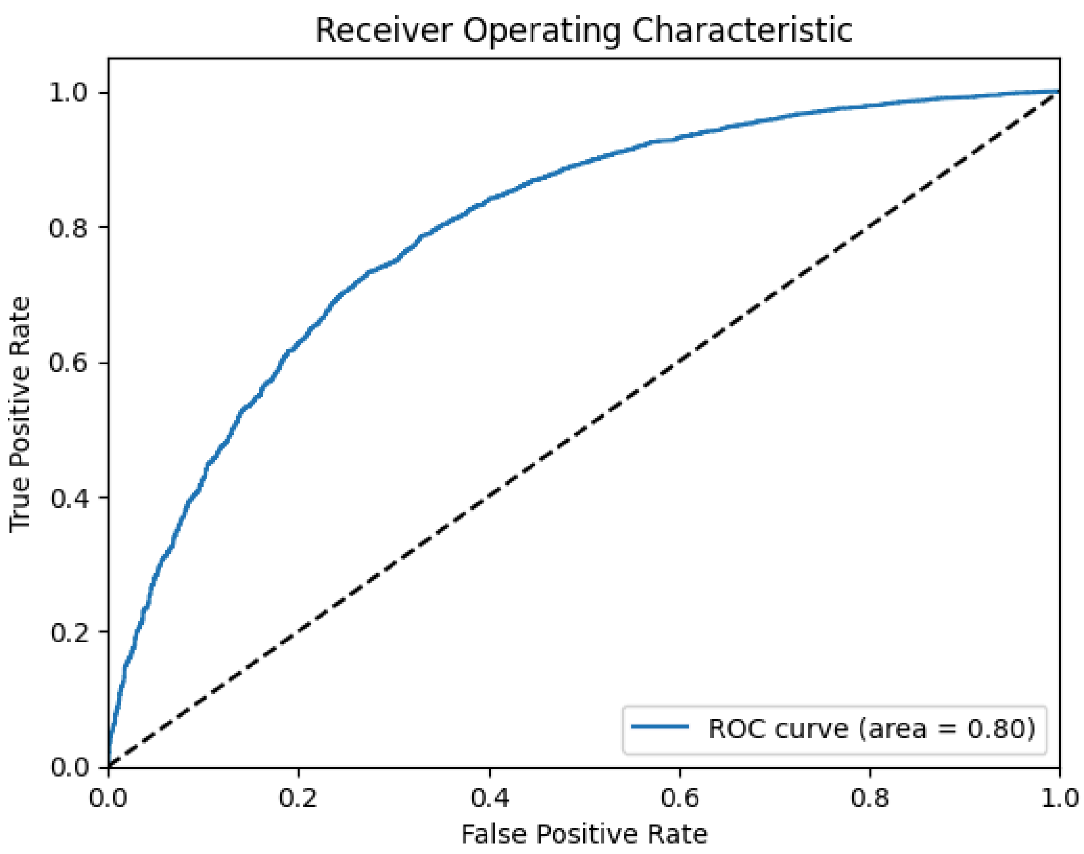


$$minimum\ average\ RMSE\ =\ 1.40386240191433$$

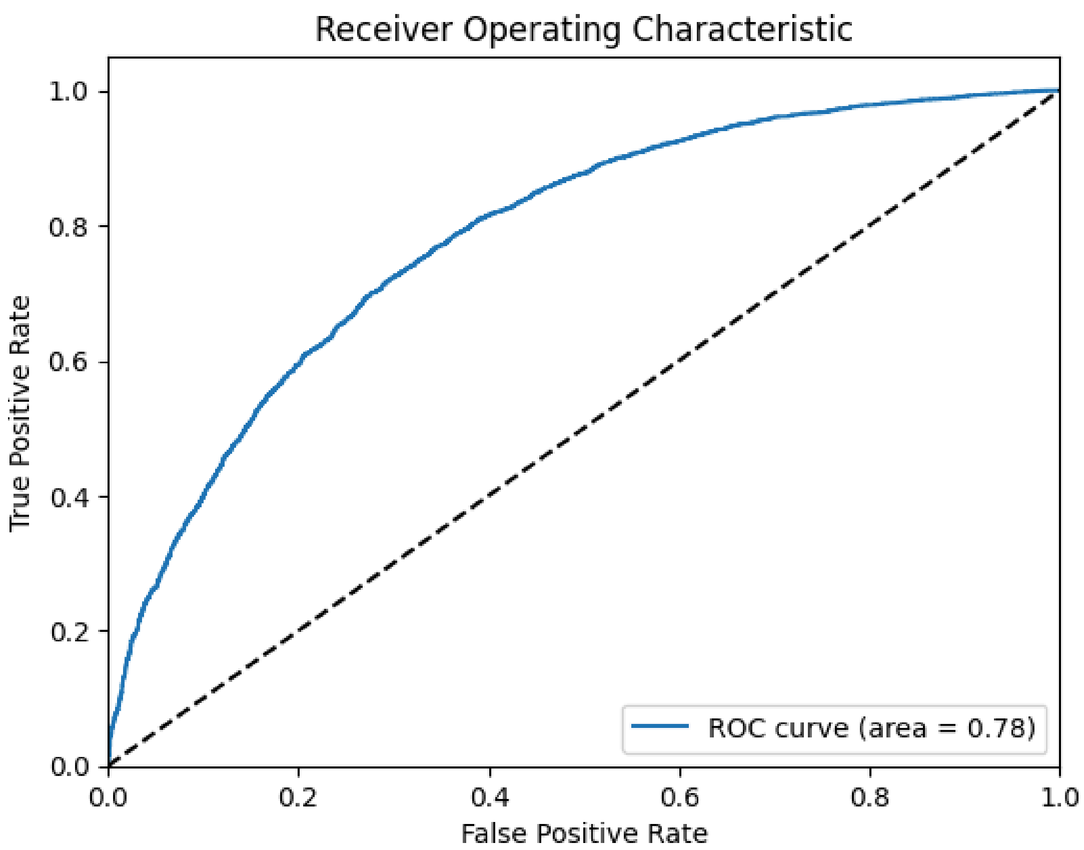
Threshold = 2.5



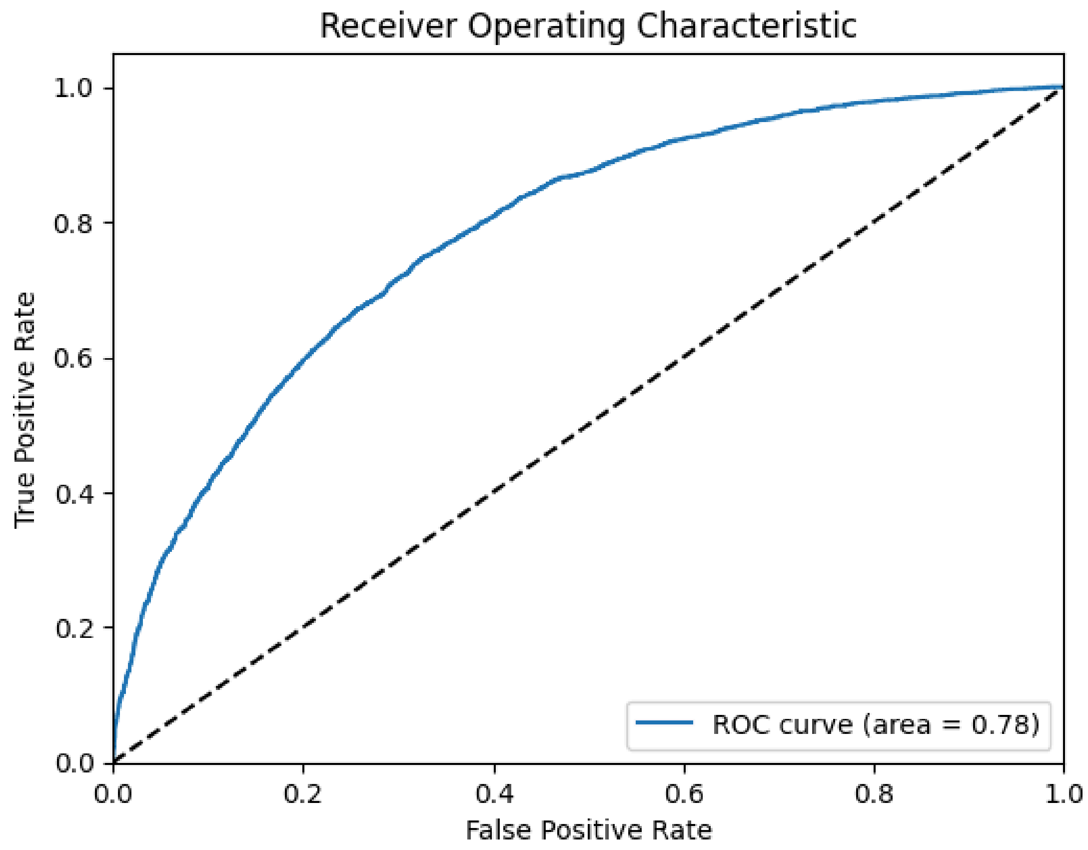
Threshold = 3.0



Threshold = 3.5



Threshold = 4.0



Question 11

$$average\ RMSE = 1.0424923612421497$$

Popular

$$average\ RMSE = 1.0355497670416107$$

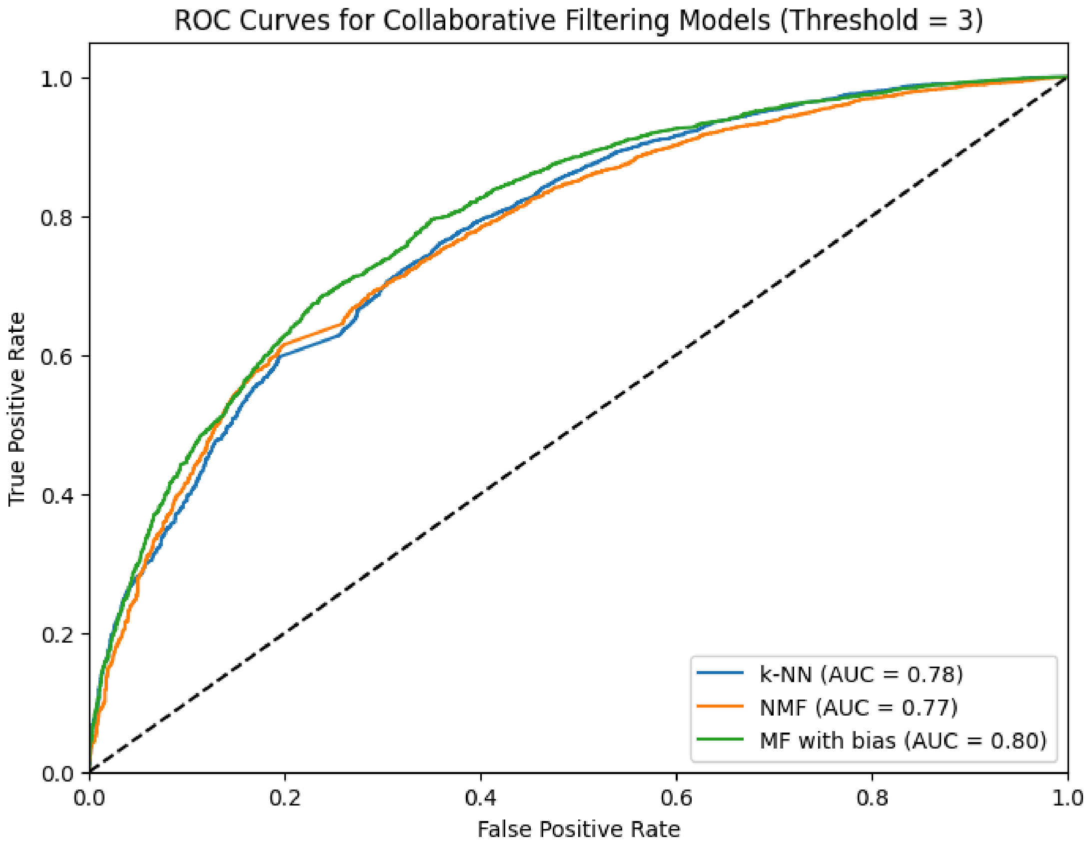
Unpopular

$$average\ RMSE = 1.1082375281838452$$

High Variance

$$average\ RMSE = 1.510530220855013$$

Question 12



The performance of the MF with bias is the best.

Question 13

```
Total number of unique queries: 50000
Distribution of relevance labels:
0    0.520136
1    0.321849
2    0.132855
3    0.017761
4    0.007400
```

Question 14

```
Fold 1:
nDCG@3: 0.5307212739772434, nDCG@5: 0.6608397947263838, nDCG@10:
0.7480978396024439
Fold 2:
nDCG@3: 1.0, nDCG@5: 0.9999999999999999, nDCG@10:
0.933745776545611
Fold 3:
nDCG@3: 0.9413401592471554, nDCG@5: 0.8845707994194446, nDCG@10:
0.8725071889285467
Fold 4:
nDCG@3: 0.9413401592471554, nDCG@5: 0.9576049743407978, nDCG@10:
0.9551307237285457
Fold 5:
nDCG@3: 0.765360636988621 8, nDCG@5: 0.8304198973631918, nDCG@10:
0.8899541168509599
```

Question 15

```
Fold 1:
Top 5 most important features:
         Feature    Importance
133  feature_134  23829.122202
7      feature_8   4256.551221
54    feature_55   4055.480095
107  feature_108   4049.734442
129  feature_130   3655.614255
Fold 2:
Top 5 most important features:
         Feature    Importance
133  feature_134  23587.659372
7      feature_8   5133.581032
54    feature_55   4366.728317
107  feature_108   4143.336742
129  feature_130   4079.324119
Fold 3:
Top 5 most important features:
         Feature    Importance
133  feature_134  23211.959232
54    feature_55   4998.220501
107  feature_108   4193.361015
129  feature_130   4028.027842
7      feature_8   3690.110570
Fold 4:
Top 5 most important features:
         Feature    Importance
133  feature_134  23760.985505
7      feature_8   4632.884738
54    feature_55   3899.246536
129  feature_130   3349.486992
128  feature_129   3220.559216
Fold 5:
Top 5 most important features:
         Feature    Importance
133  feature_134  23480.303283
7      feature_8   4791.326552
54    feature_55   4058.867884
107  feature_108   3495.305341
129  feature_130   3188.522764
```

Question 16

```
Fold 1:
Reduced: nDCG@3: 0.5307212739772434, nDCG@5: 0.6608397947263838, nDCG@10: 0.7799082337019198
Reduced 60: nDCG@3: 0.6173196815056892, nDCG@5: 0.5557046000229097, nDCG@10: 0.5914468270998005
Fold 2:
Reduced: nDCG@3: 1.0, nDCG@5: 0.9634829125393233, nDCG@10: 0.9126821207086606
Reduced 60: nDCG@3: 0.5586598407528446, nDCG@5: 0.542395025659202, nDCG@10: 0.5471110380319288
Fold 3:
Reduced: nDCG@3: 0.8826803184943108, nDCG@5: 0.8824086793035104, nDCG@10: 0.9236912790150748
```

Reduced 60: nDCG@3: 0.6173196815056892, nDCG@5: 0.6175913206964894, nDCG@10: 0.580138981986594
Fold 4:
Reduced: nDCG@3: 0.9413401592471554, nDCG@5: 0.9576049743407978, nDCG@10: 0.97248852921274
Reduced 60: nDCG@3: 0.6173196815056892, nDCG@5: 0.654108408157166, nDCG@10: 0.6818128270805792
Fold 5:
Reduced: nDCG@3: 1.0, nDCG@5: 0.9999999999999999, nDCG@10: 0.943718471705651
Reduced 60: nDCG@3: 0.75, nDCG@5: 0.7007980959328717, nDCG@10: 0.6703309994925489