Large-Scale Data Mining: Models and Algorithms ECE 219 MSOL Spring 2022

Prof. Vwani Roychowdhury

UCLA, Department of ECE

Team Members:
Rohan Mehta, UID: 205871841

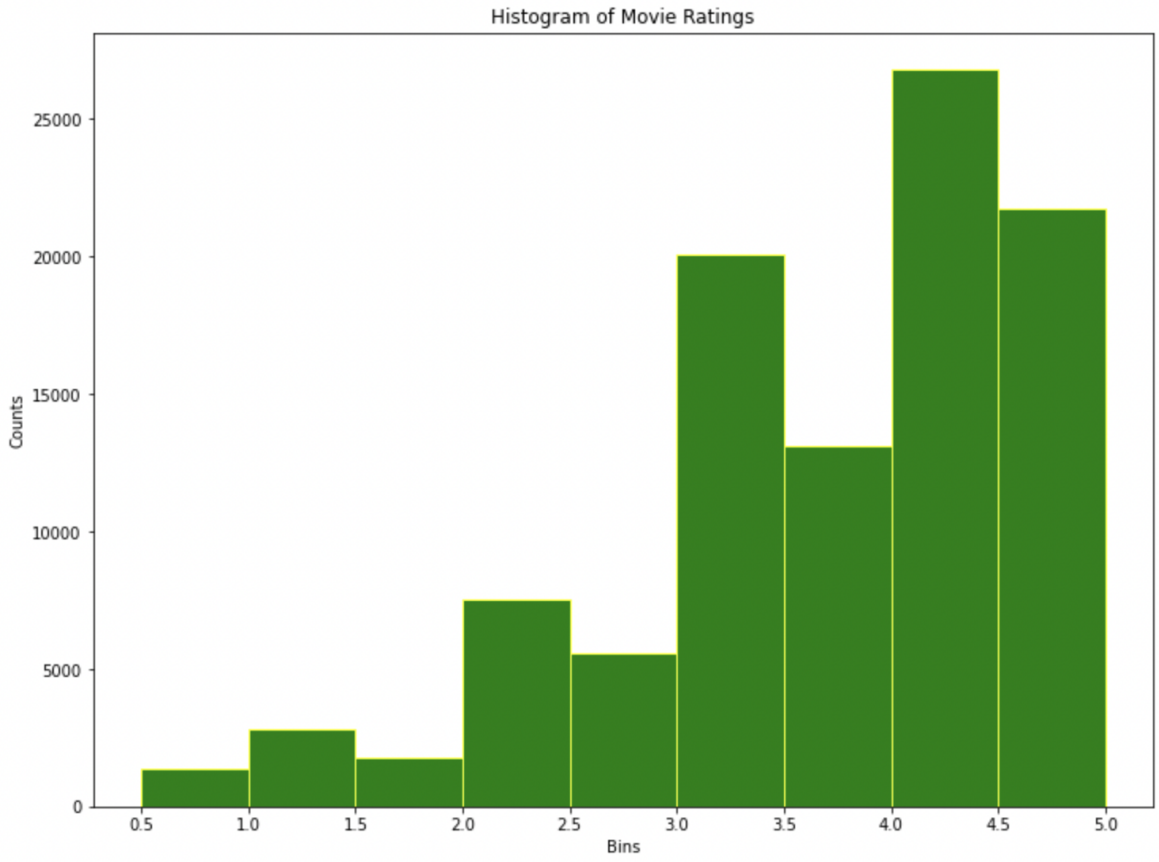
Project 3: Recommender Systems

1. Exploring the Dataset

Question 1:

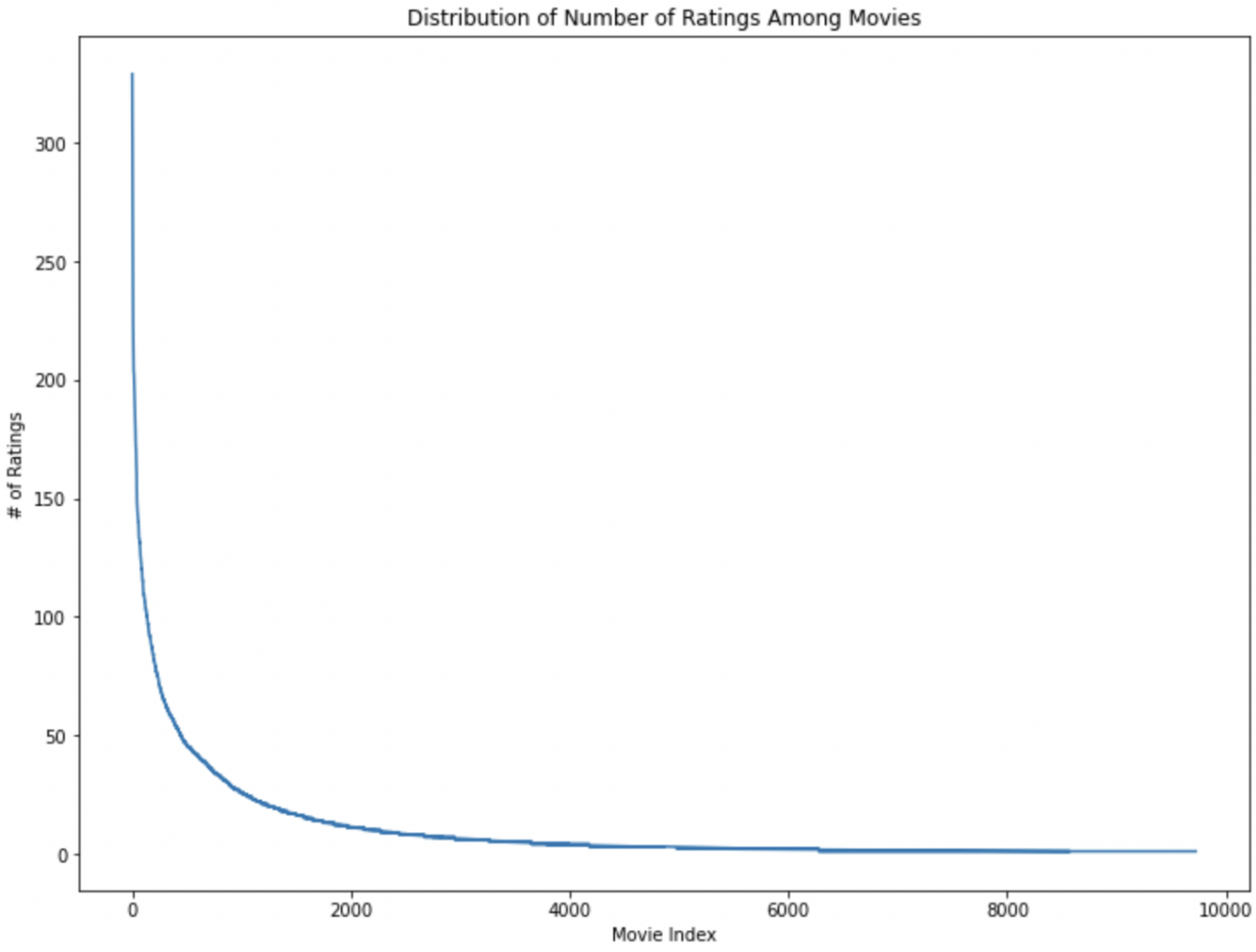
A. Sparsity = Total number of available ratings / Total number of possible ratings = **0.0169996830556**

B.

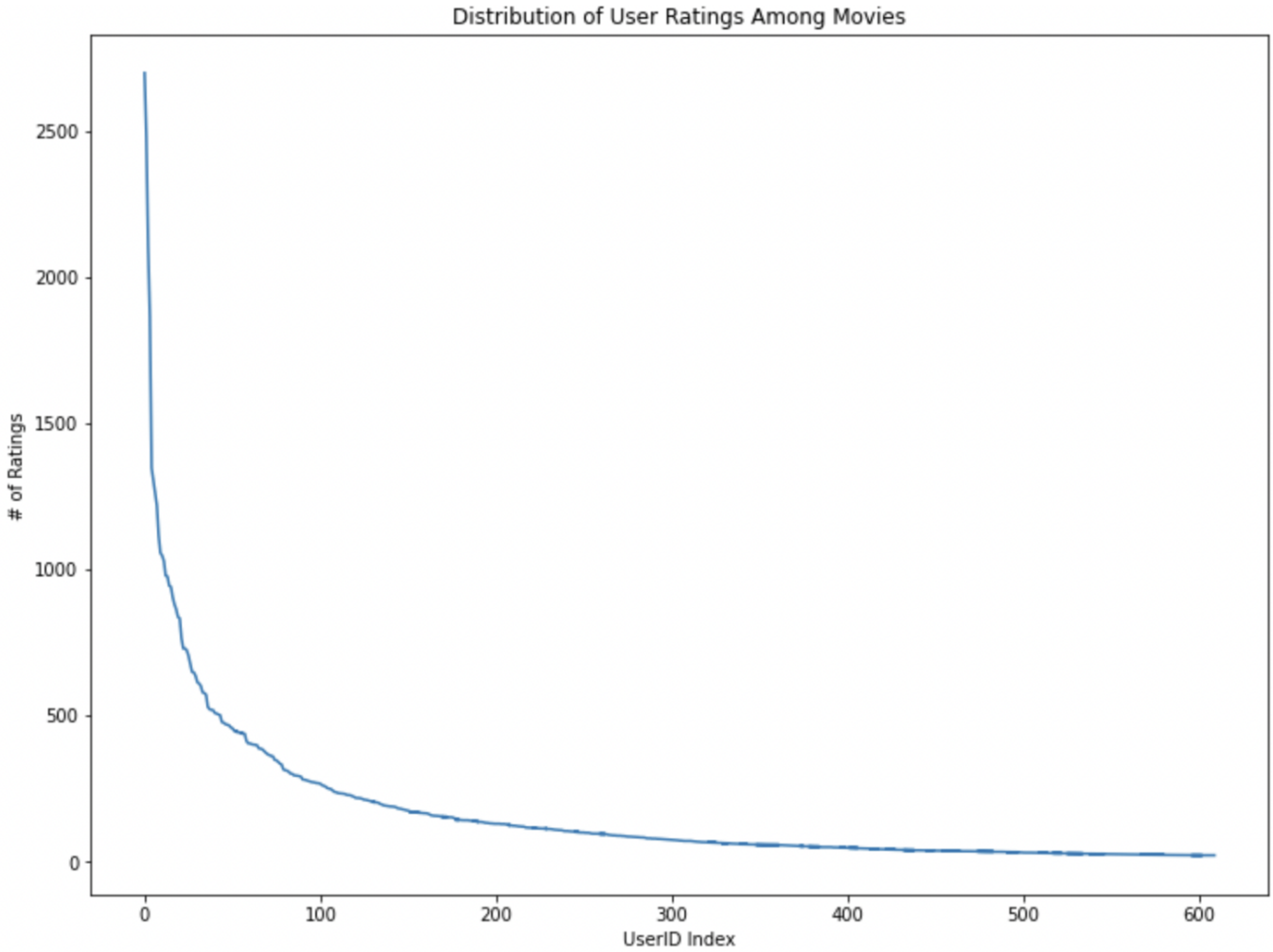


In general, as the value of the rating increases then so does the number of ratings of the movies. This suggests that users are more likely to rate a movie if they found it a pleasant experience, but less likely if they had a negative experience. This could be due to the fact that a lot of users might not finish a movie they dislike on a streaming platform, or perhaps due to other psychological biases.

C.

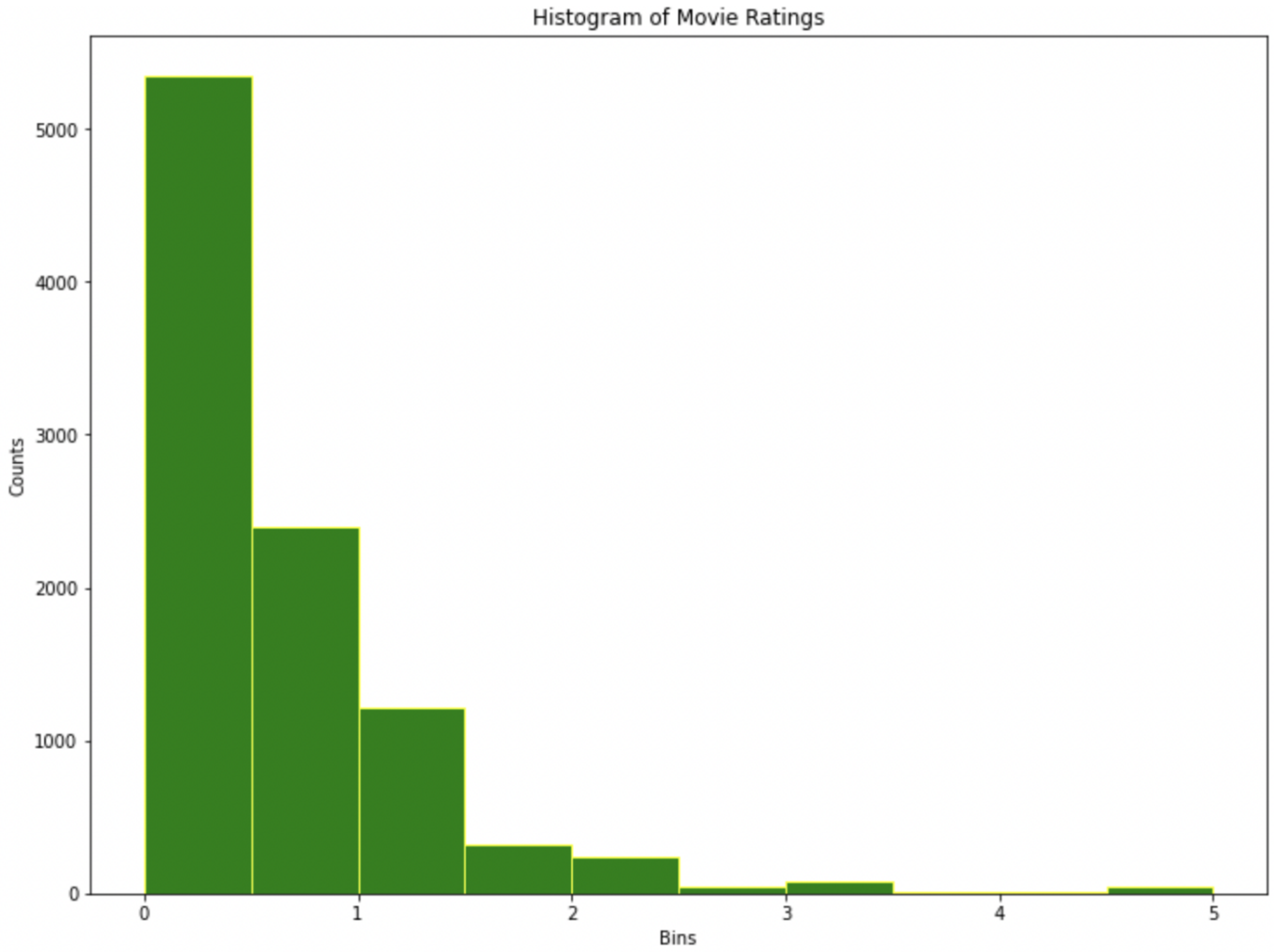


D.



E. In part C, a few movies appear to be very popular and rated by many users. For a recommendation system, these will be more likely to be predicted for a higher number of users due to its high frequency in the ratings matrix. In part D, a similar trend is present where a few users rate many movies. The ratings of these few users will have more impact on the recommendation process since they provide the majority of the data in a sparse matrix.

F.



This result suggests that bad movies are disliked to various degrees, and perhaps for a variety of reasons, whereas movies that are rated highly meet some ideal experience threshold for the general viewer. The histogram is right skewed but not monotonically decreasing.

Question 2:

A. μ_u: Mean rating for user u computed using her specified ratings

$$\frac{\sum_{k=1}^{n} r_{uk}}{n} = u_u$$

$k \in I_u$

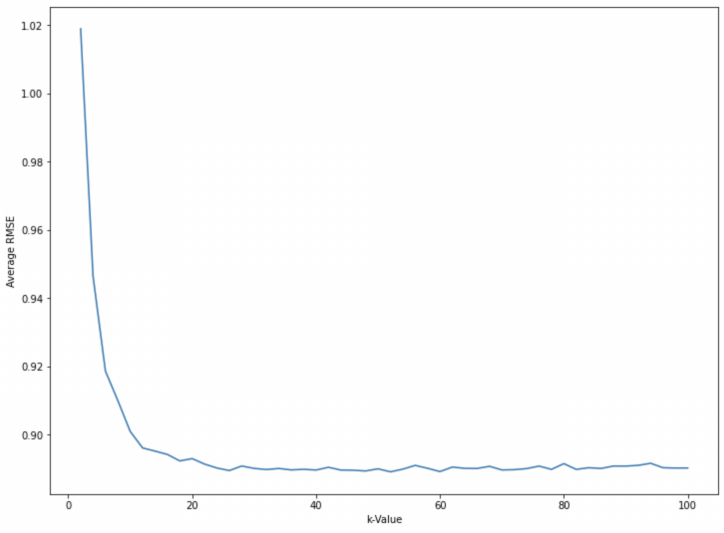
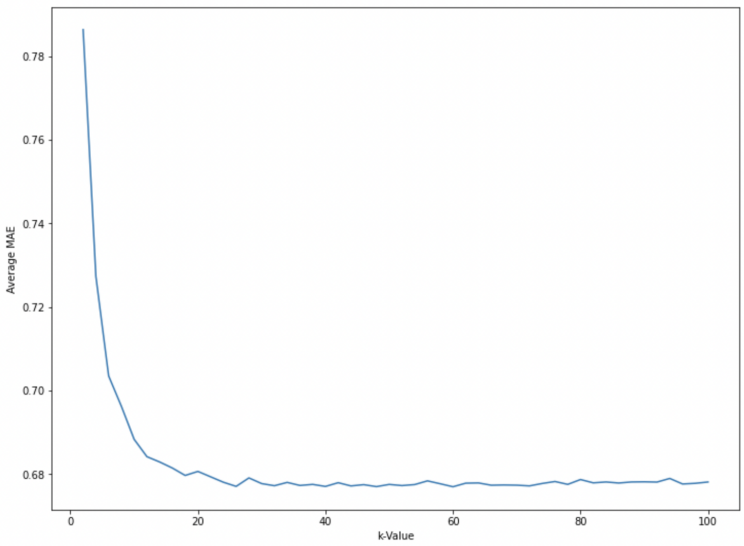
B.

The intersection of Iu and Iv denotes the set of movie indices which were rated by both users *u* and *v.* The intersection of these sets can be the null set because it is not necessary that both users selected the same movies to rate.

Question 3:

Mean centering the ratings in the prediction function helps to remove bias in the individual rating style of the users. For example, a low-rating user *u* and high-rating user *v,* may overlap in rated movies but have a low correlation coefficient due to individual rating style. Another user who rates very highly, *w,* may attain a higher correlation coefficient with user *v* even though their preferences might not be as closely tied. Mean centering helps avoid this bias issue and ultimately compensates for critical or generous raters.

Question 4:

RMSE	MAE
	

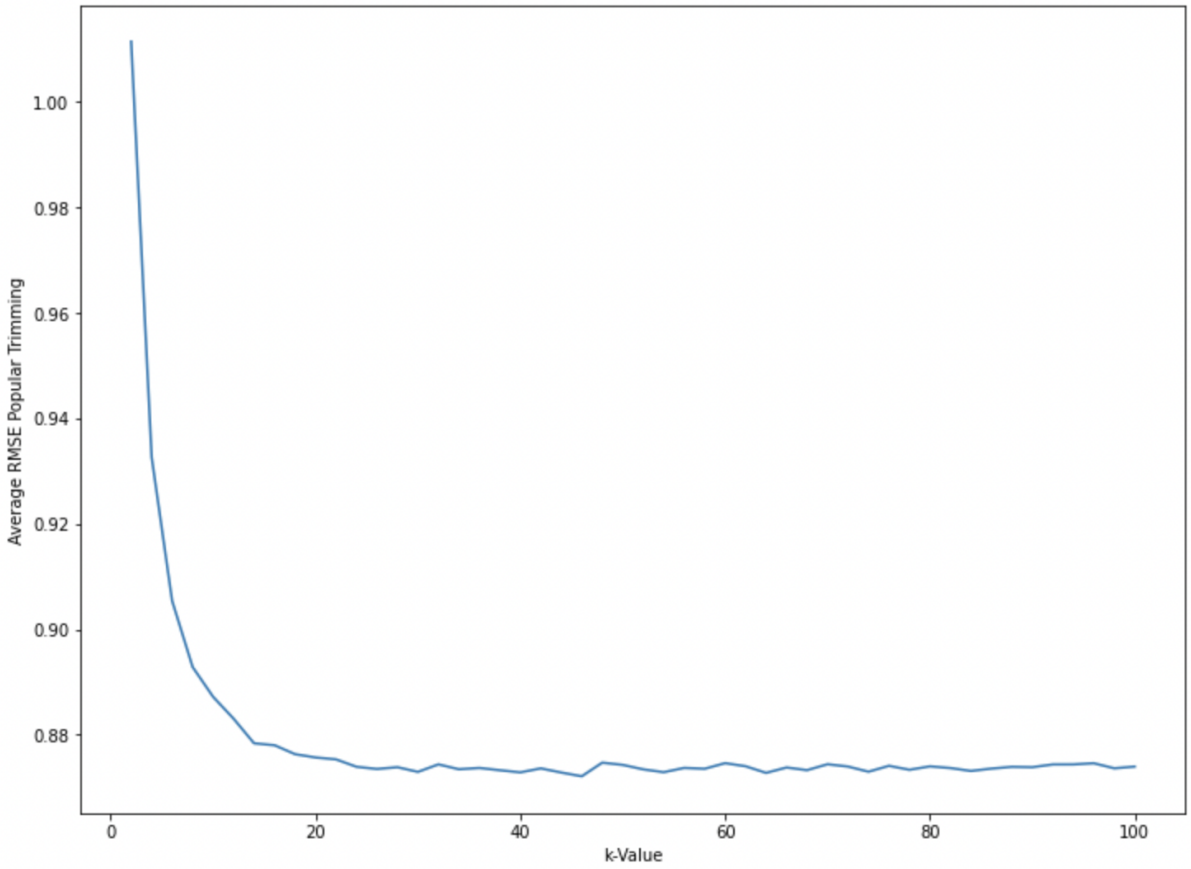
Question 5:

I used a threshold to track whether between subsequent k-values there was a significant drop in the average RMSE or MAE value. The value for k for which this threshold was triggered is 42.

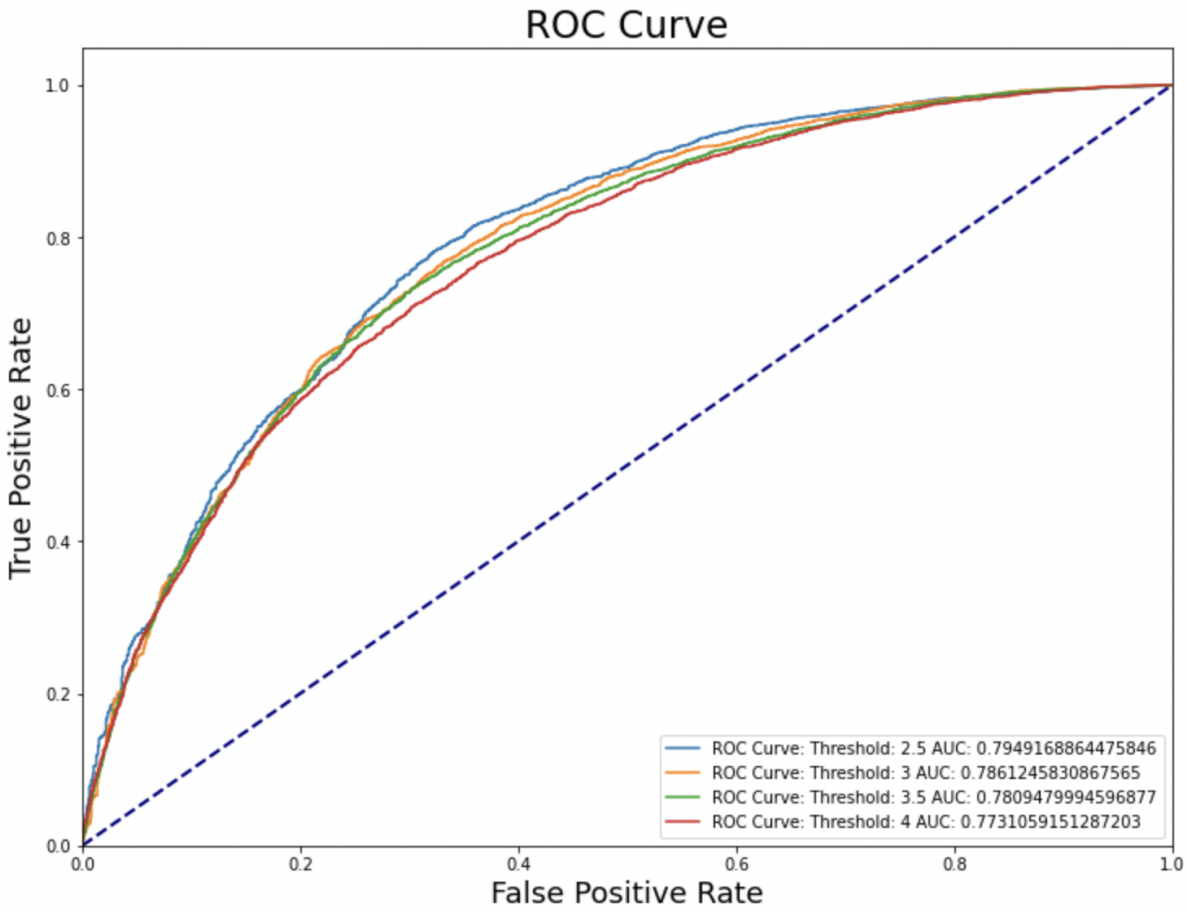
Avg. RMSE, k = 42	Avg. MAE, k = 42
0.88955	0.67715

Question 6:

Popular Trimming

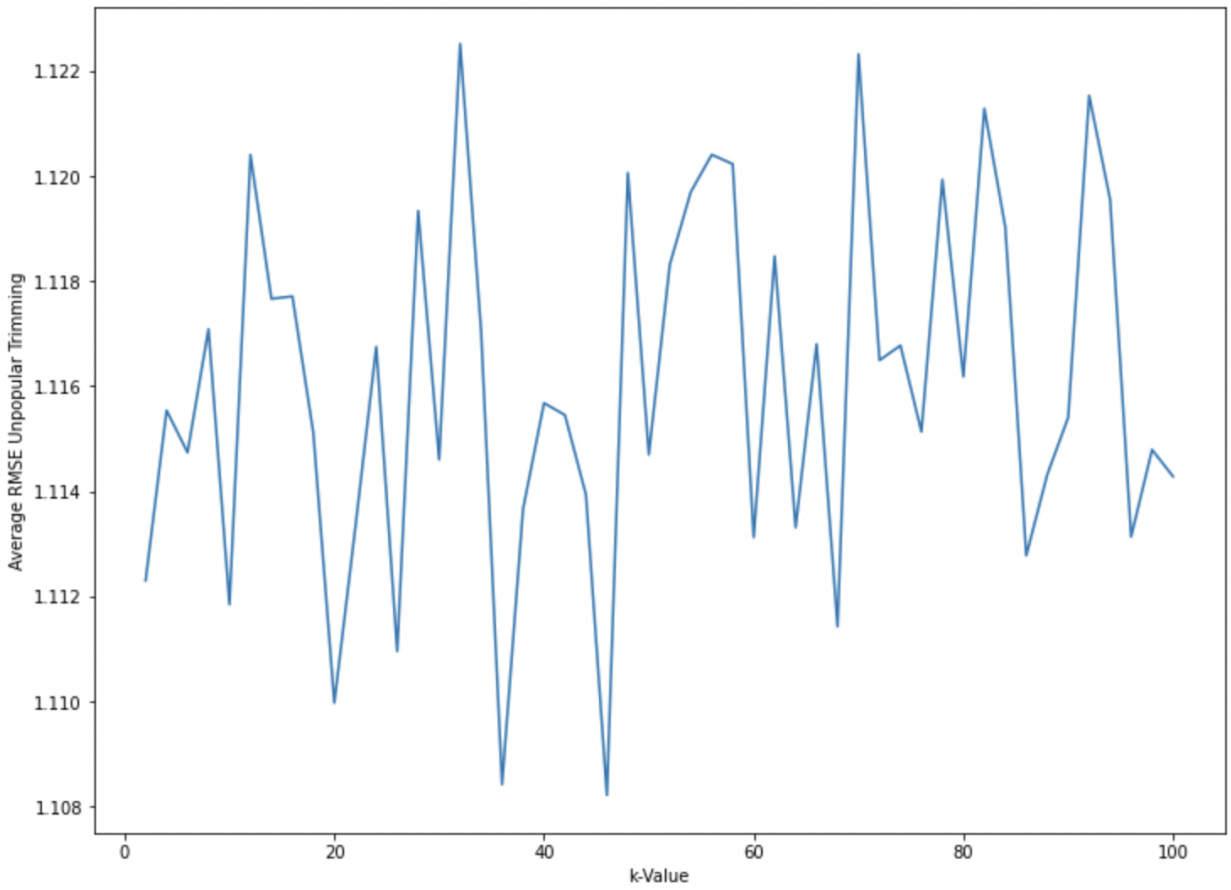


The minimum average RMSE value associated with the popular trimming 0.87213748

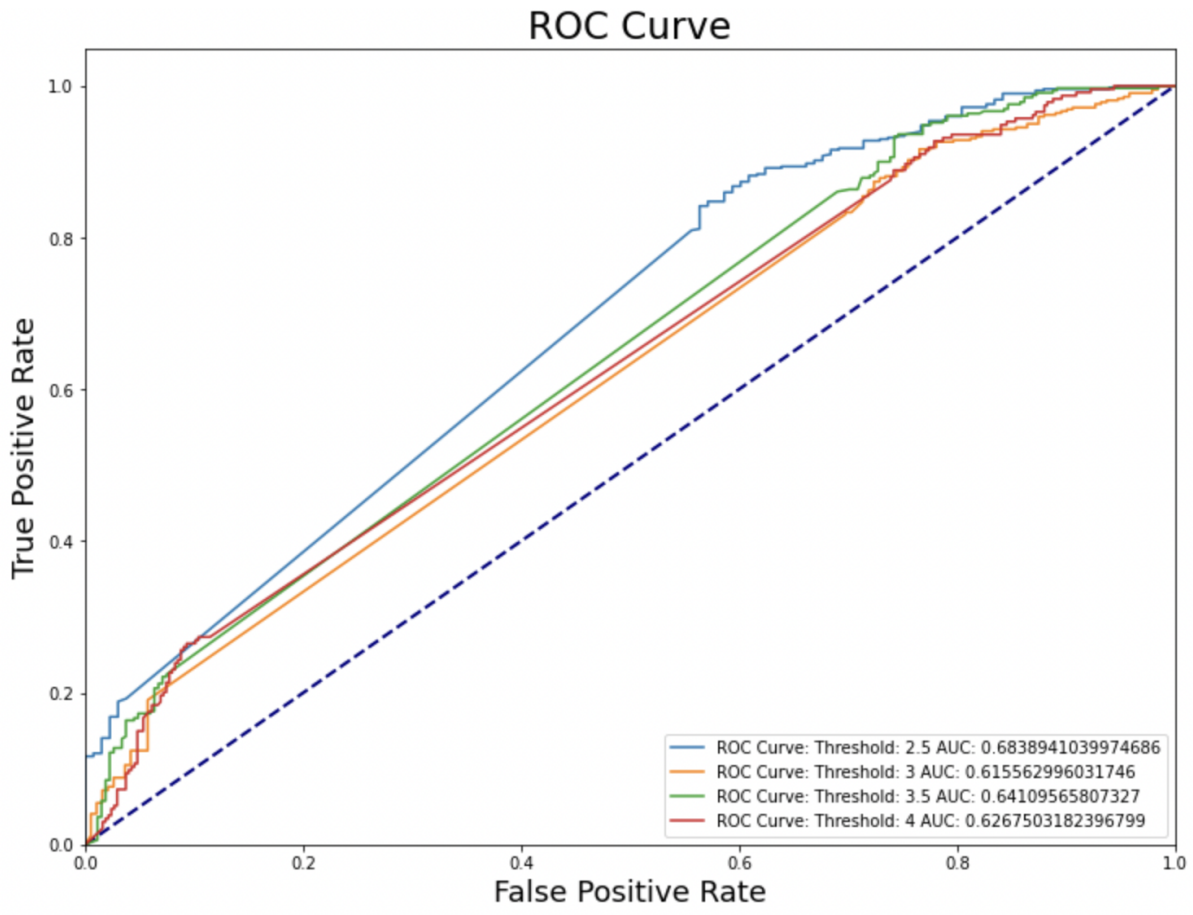


Ratings Threshold	AUC
2.5	0.7949
3.0	0.7861
3.5	0.7809
4.0	0.7731

Unpopular Trimming

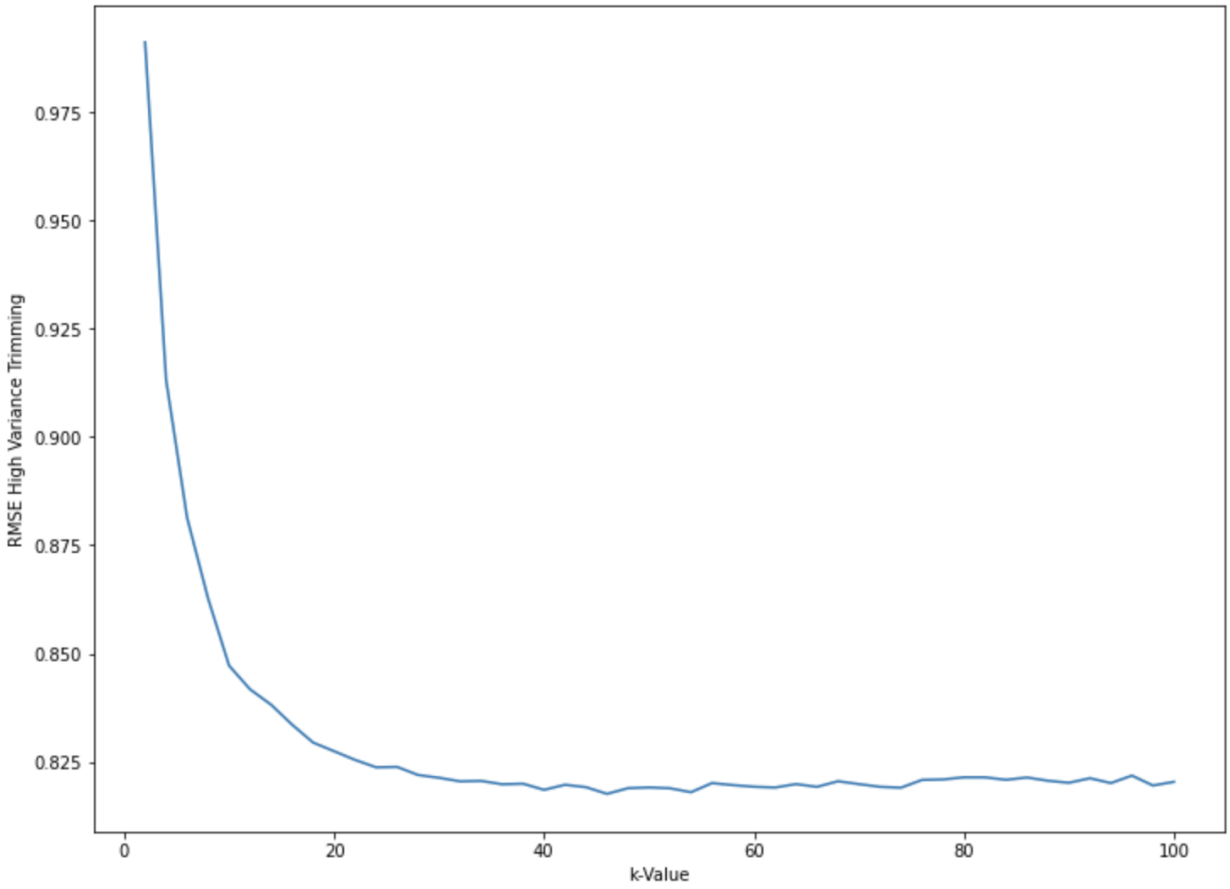


The minimum average RMSE value associated with the unpopular trimming 1.10821028

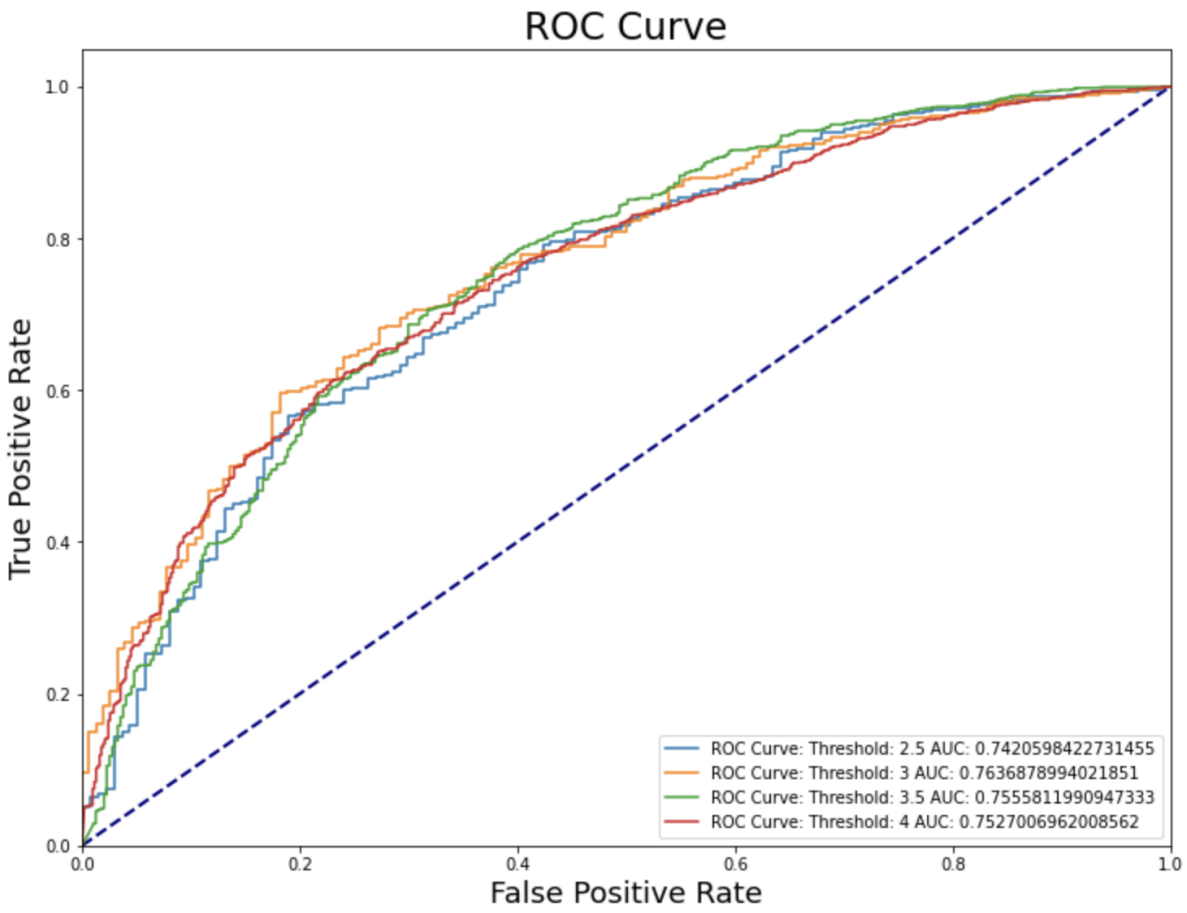


Ratings Threshold	AUC
2.5	0.68389
3.0	0.61556
3.5	0.64110
4.0	0.62675

High Variance Trimming

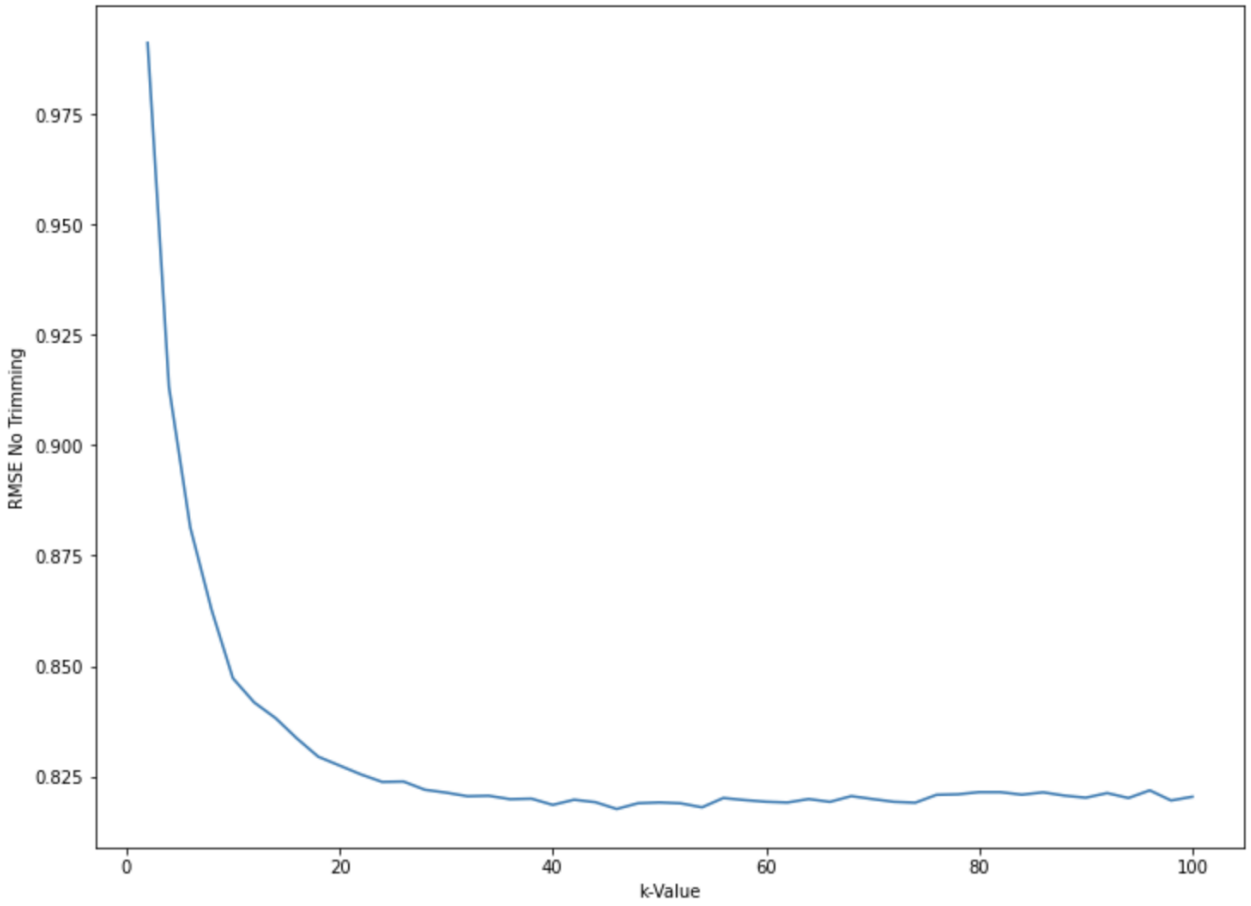


The minimum average RMSE value associated with high variance trimming was 0.8176602

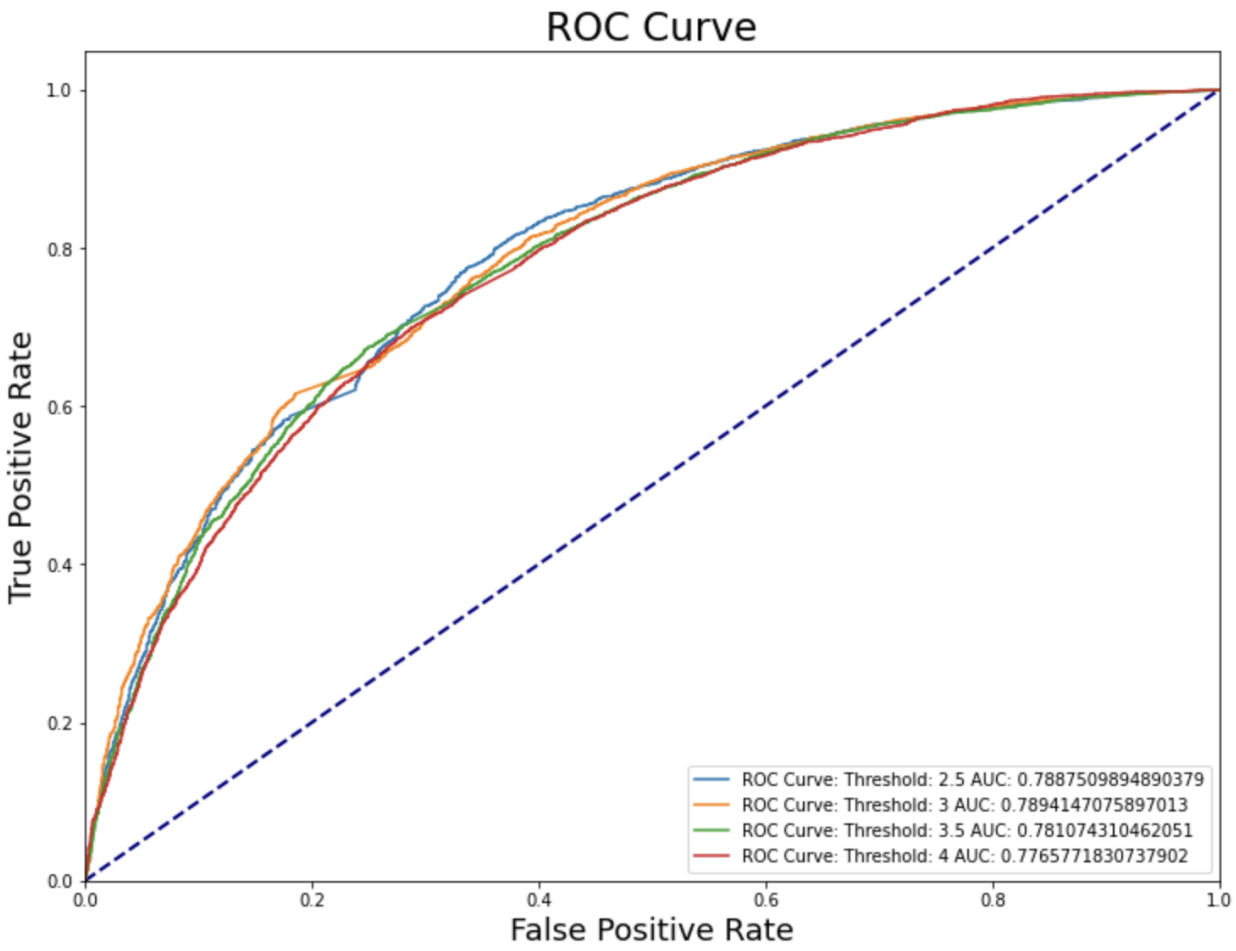


Ratings Threshold	AUC
2.5	0.74206
3.0	0.76369
3.5	0.75558
4.0	0.75270

No Trimming



The minimum average RMSE value associated with no trimming was 0.888951



Ratings Threshold	AUC
2.5	0.78875
3.0	0.78941
3.5	0.78107
4.0	0.77658

Question 7:

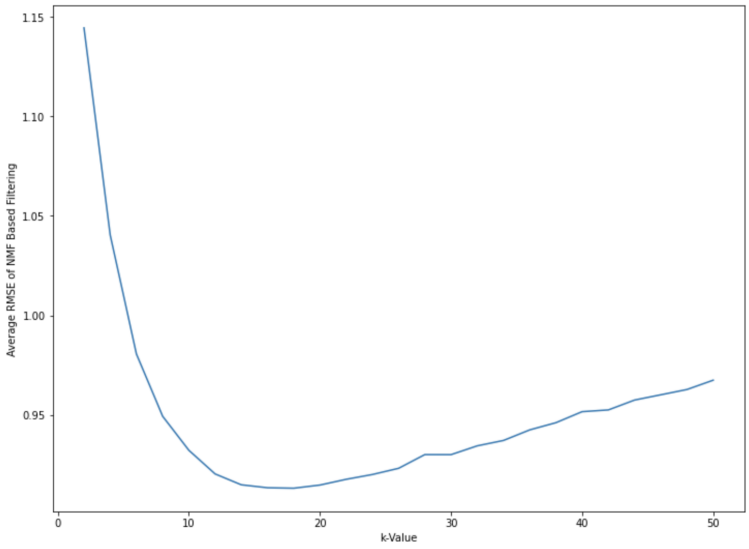
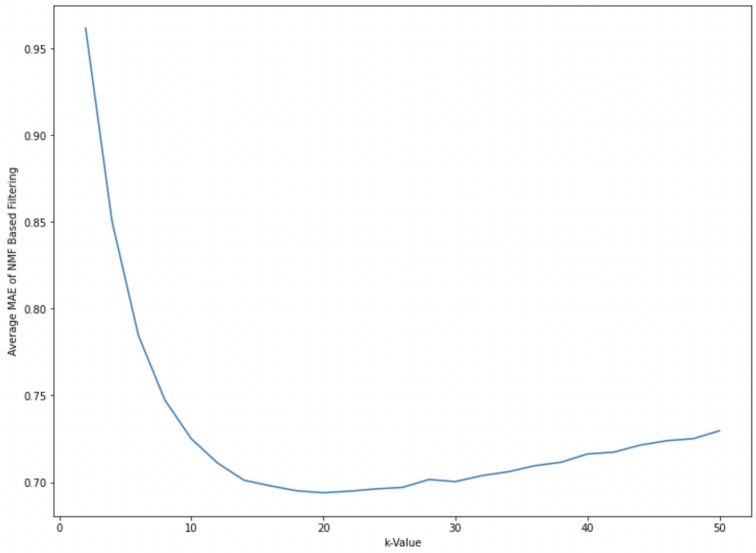
$$\underset{U,V}{\text{minimize}} \quad \sum_{i=1}^{m}\sum_{j=1}^{n} W_{ij}(r_{ij} - (UV^T)_{ij})^2 \qquad (5)$$

The optimization problem given by equation 5 above is not convex because it has more than one global minima.
When fixing U in the equation above, one can turn the problem into a least-squares problem in the following way:

$$V = (UU^T + \lambda I)^{-1} UR$$

Question 8:

A.

RMSE	MAE
	

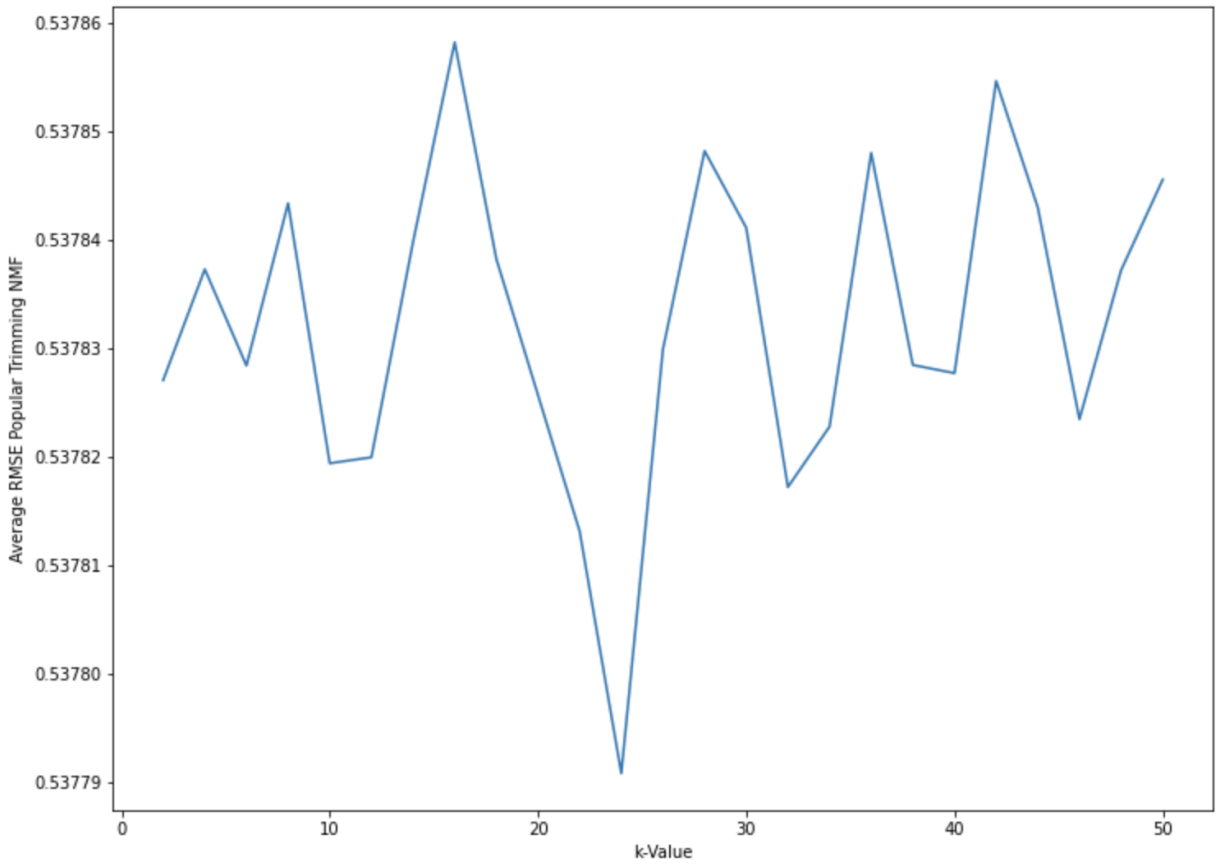
B.

	RMSE	MAE
Optimal Latent Factors	14	22
Average Value	0.913965	0.692765

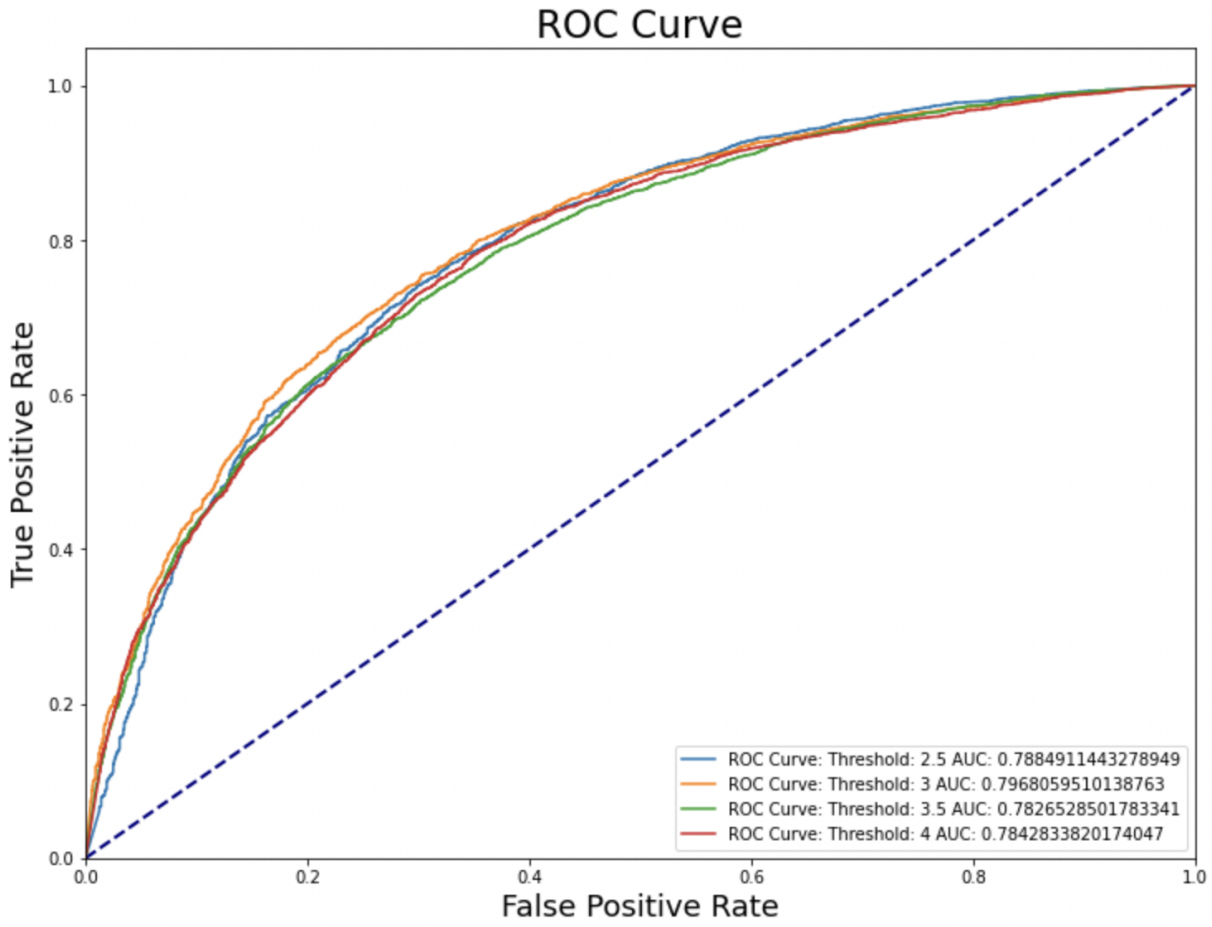
After separating the genres column from the movies.csv column, I used the re library to define a delimiter which separated the different genres. The number of genres after this process comes out to 19. The optimal number of latent factors for MAE is closer to the number of genres than RMSE.

C.

Popular Trimming

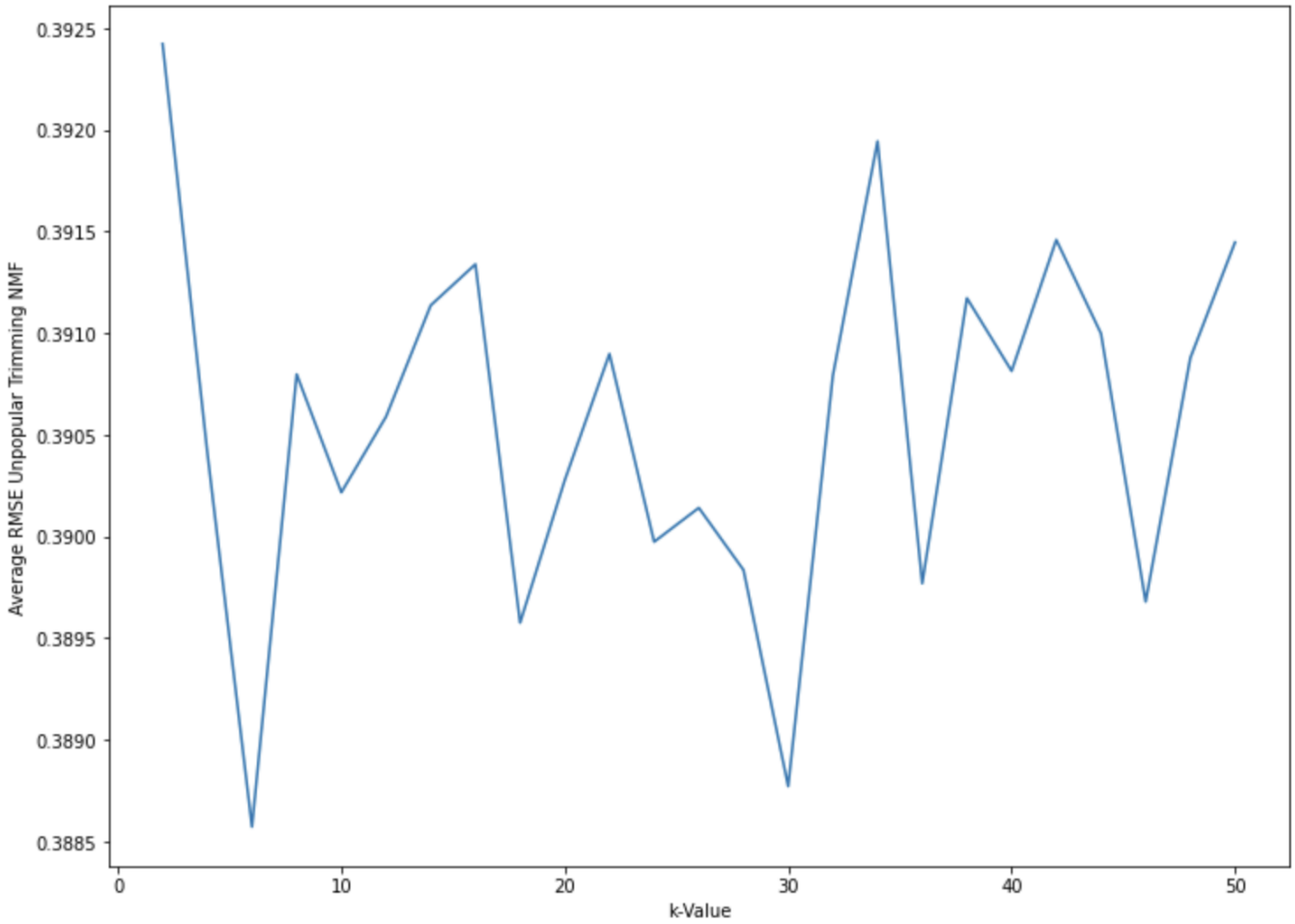


The minimum average RMSE value associated with no trimming was 0.53779

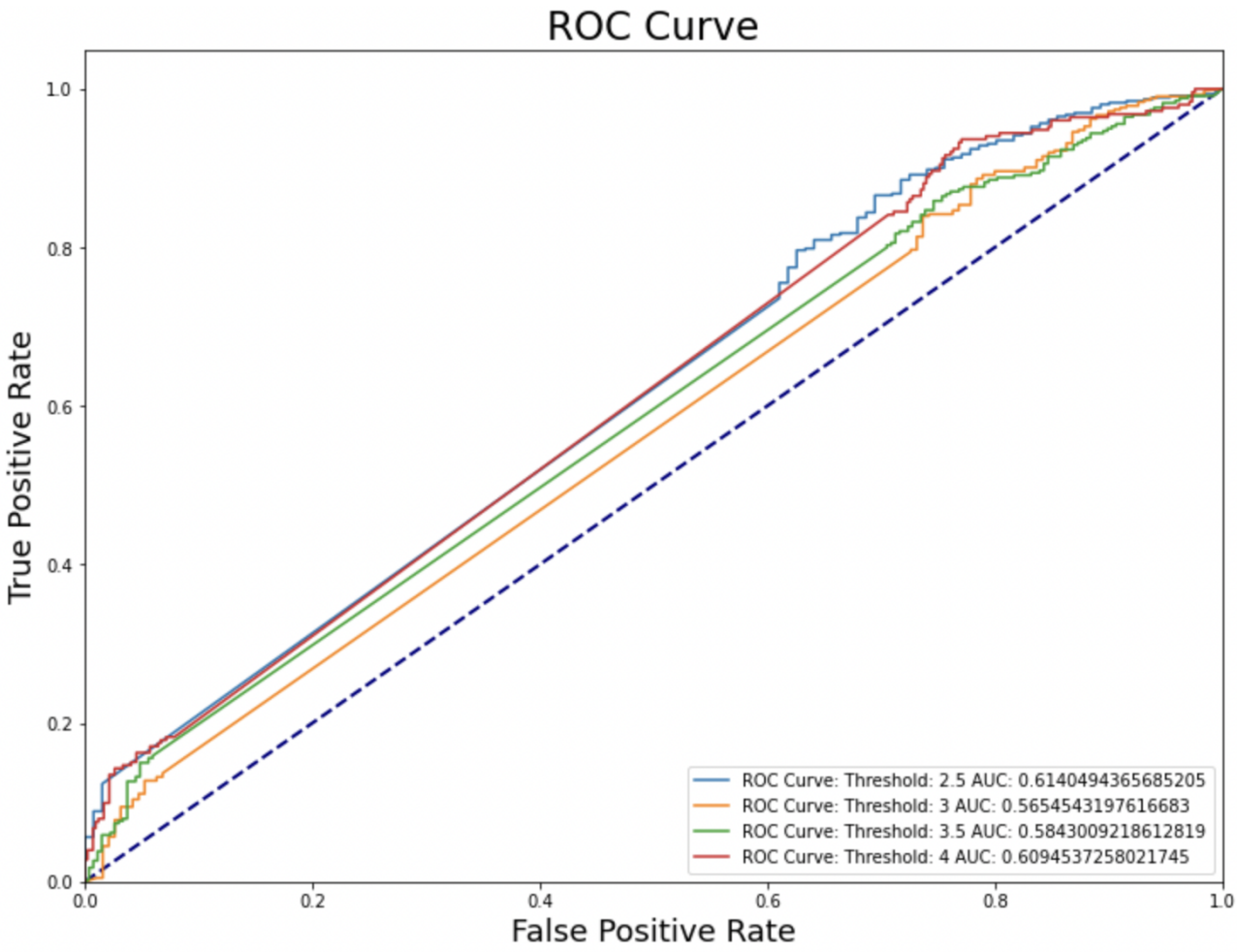


Ratings Threshold	AUC
2.5	0.78849
3.0	0.79681
3.5	0.78265
4.0	0.78428

Unpopular Trimming

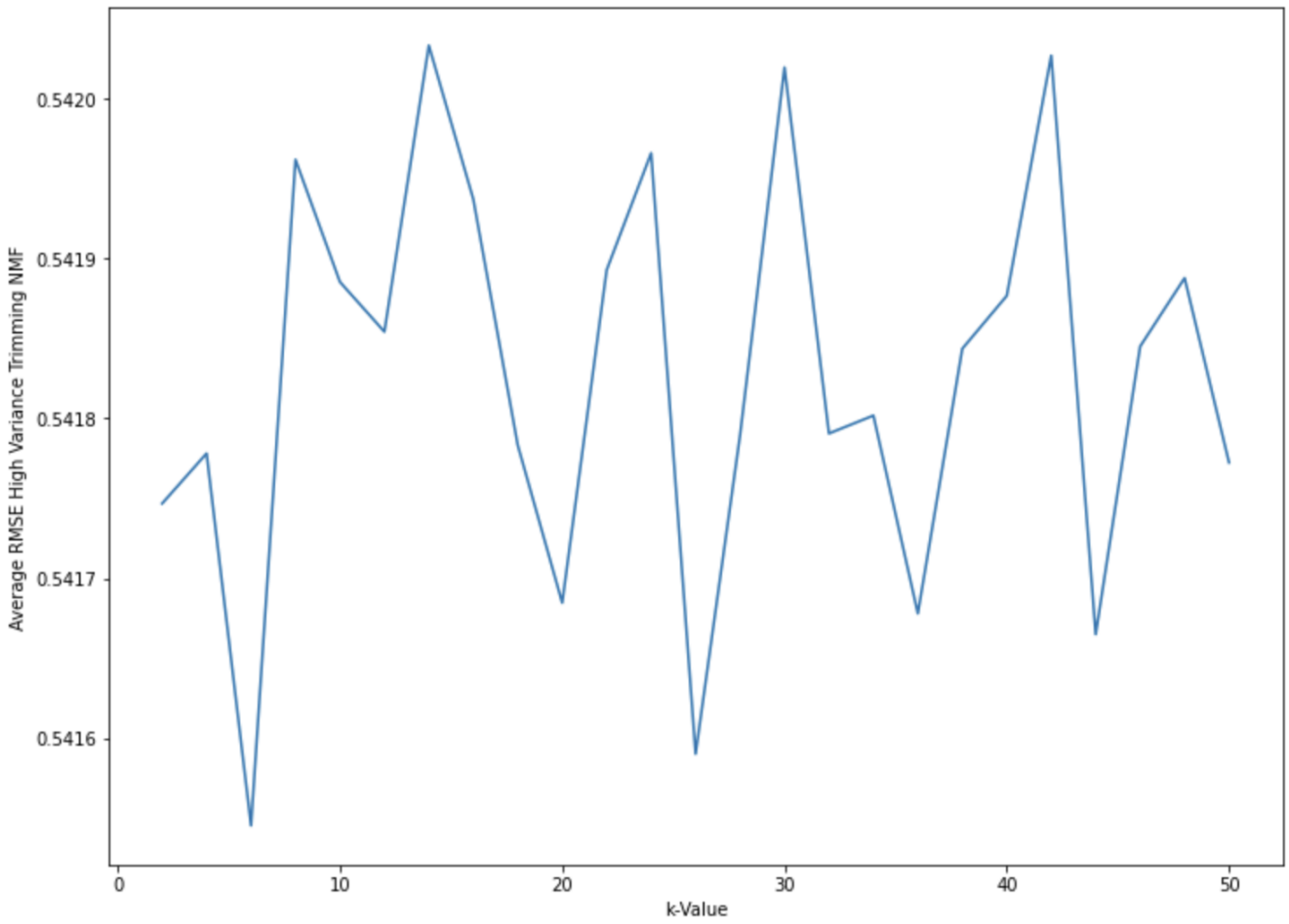


The minimum average RMSE value associated with no trimming was 0.388572

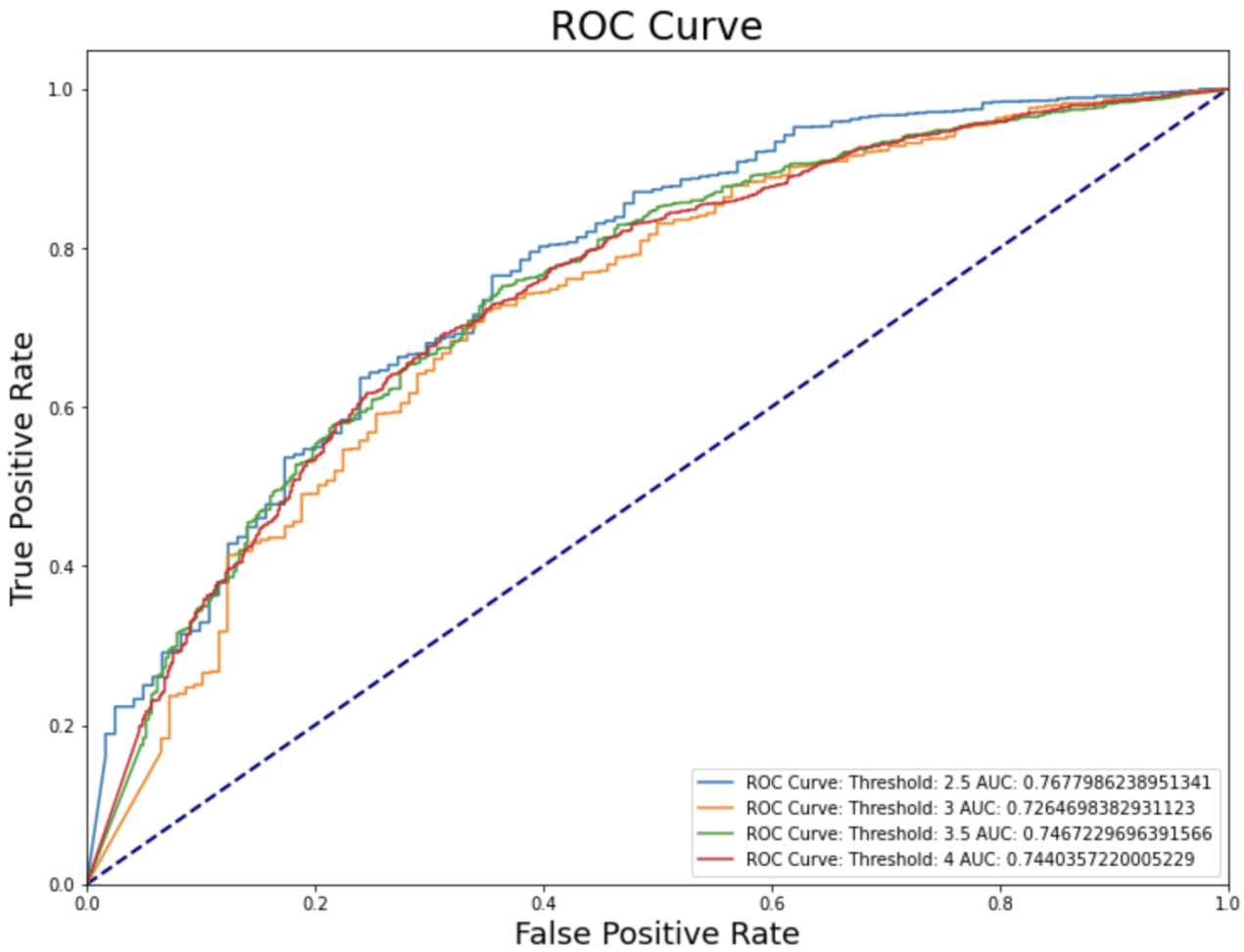


Ratings Threshold	AUC
2.5	0.614049
3.0	0.56545
3.5	0.58430
4.0	0.60945

High Variance Trimming



The minimum average RMSE value associated with no trimming was 0.54155



Ratings Threshold	AUC
2.5	0.76779
3.0	0.72646
3.5	0.74672
4.0	0.75504

Question 9:

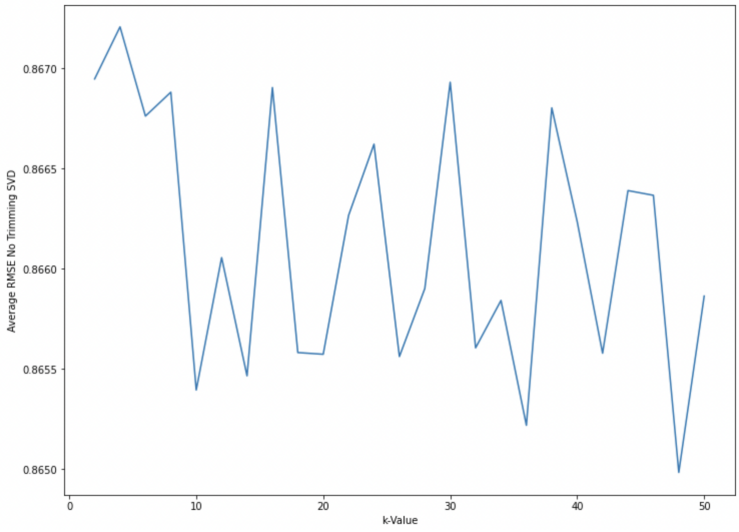
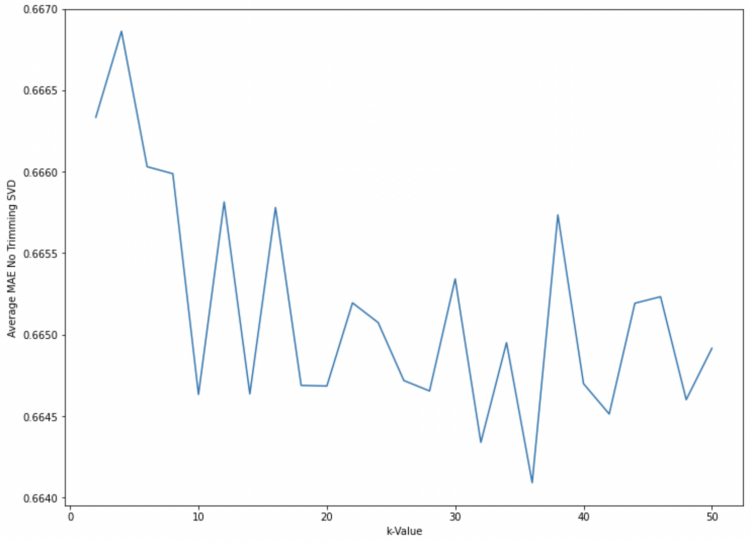
Genres of Top Ten Movies in *V*		
Column 1	**Column 4**	**Column 19**
Comedy|Drama|Romance Adventure|Children Comedy|Crime|Drama Adventure|Animation|Children|Fantasy|Musical Comedy|Drama Drama|Fantasy|Horror|Thriller Drama Action|Adventure Action|Drama|Sci-Fi Adventure|Drama|Sci-Fi	Comedy Comedy Action|Sci-Fi|Thriller Action|Crime|Drama Comedy Adventure|Children Comedy|Drama Comedy|Drama Mystery|Thriller Drama	Drama|Romance Drama|Romance Action|Adventure|Sci-Fi|Thriller Action|Crime|Drama|Thriller Drama|Romance Action|Adventure|Drama Drama Drama Crime|Drama Drama

The top ten movies per column in *V* definitely highlight different collections of genre based on which column is being inspected. For example, Column 1 seems to favor Action and Adventure whereas Column 4 is more heavily oriented towards Comedy and Drama. Column 19 is dominated by Drama with a combination of Romance and a couple other genres.

Based on the pattern identified above, there does seem to be a connection between the latent factors and movie genres. If a user had a high preference for drama across their ratings, they would appear in the 19th column of the *U* matrix because an association would exist between that user and this particular collection of genres.

Question 10:

A.

RMSE	MAE
	

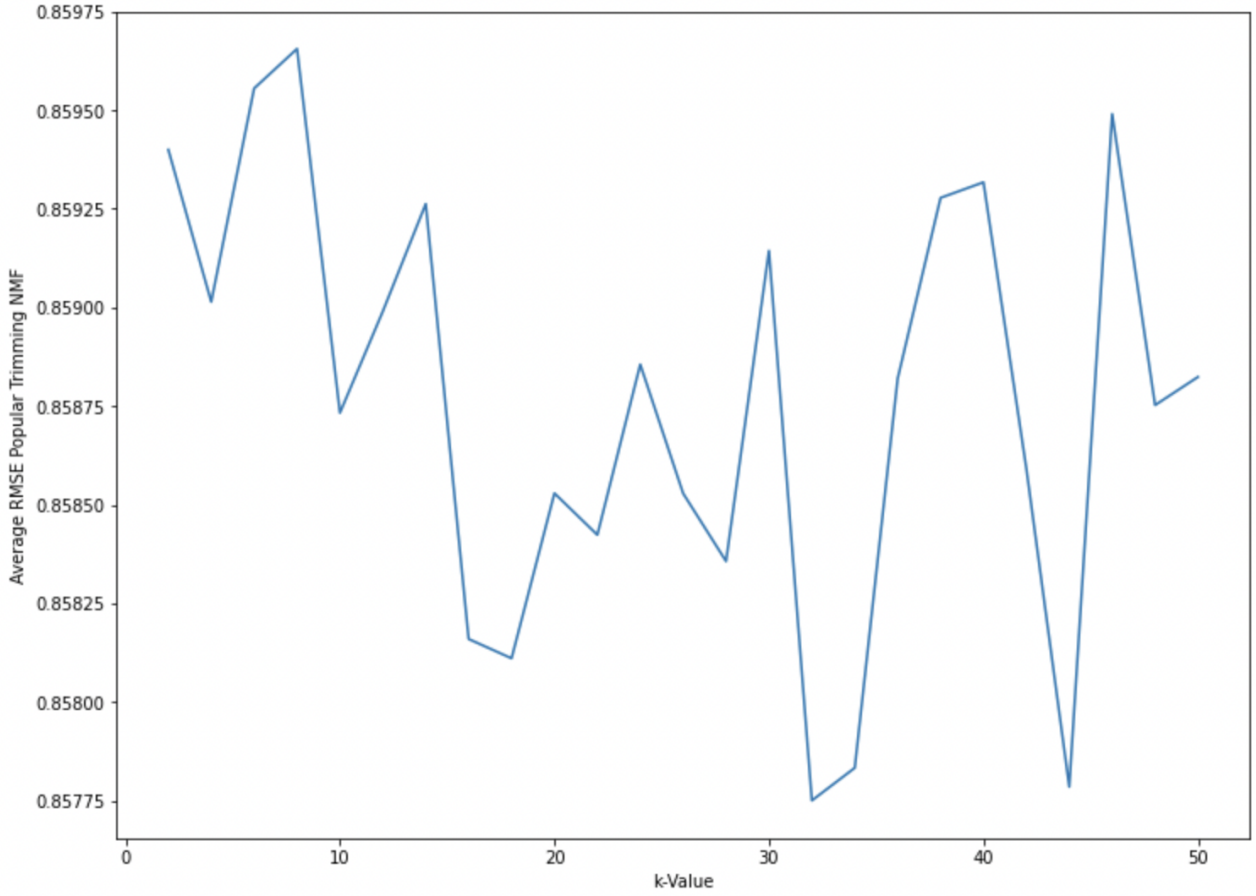
B.

	RMSE	MAE
Optimal Latent Factors	48	36
Average Value	0.86498	0.66409

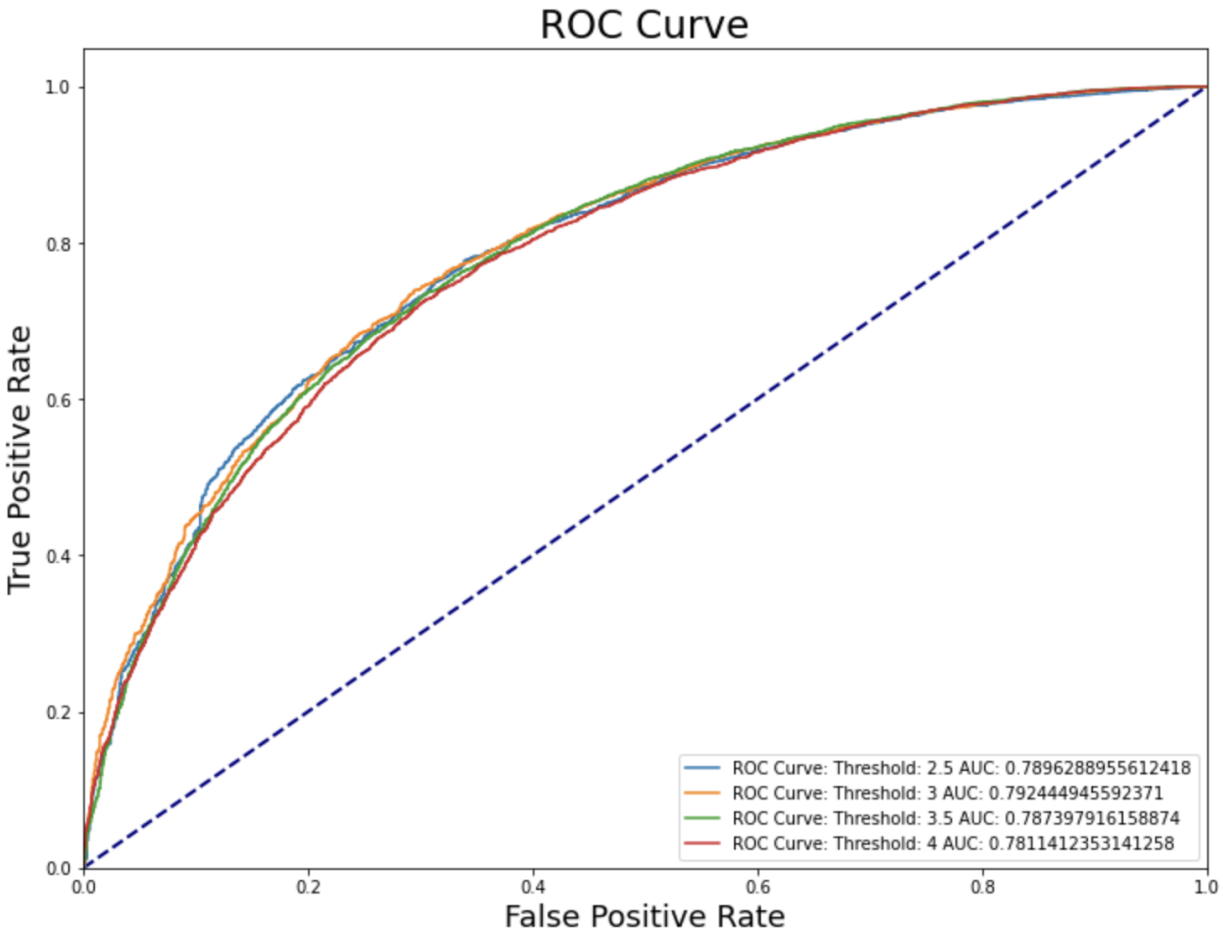
After inspecting the plots shown above, the optimal latent factors were shown to be 48 and 36 respectively for RMSE and MAE. The number of genres, as calculated earlier, is 19 which does not correspond to the optimal number of latent factors.

C.

Popular Trimming

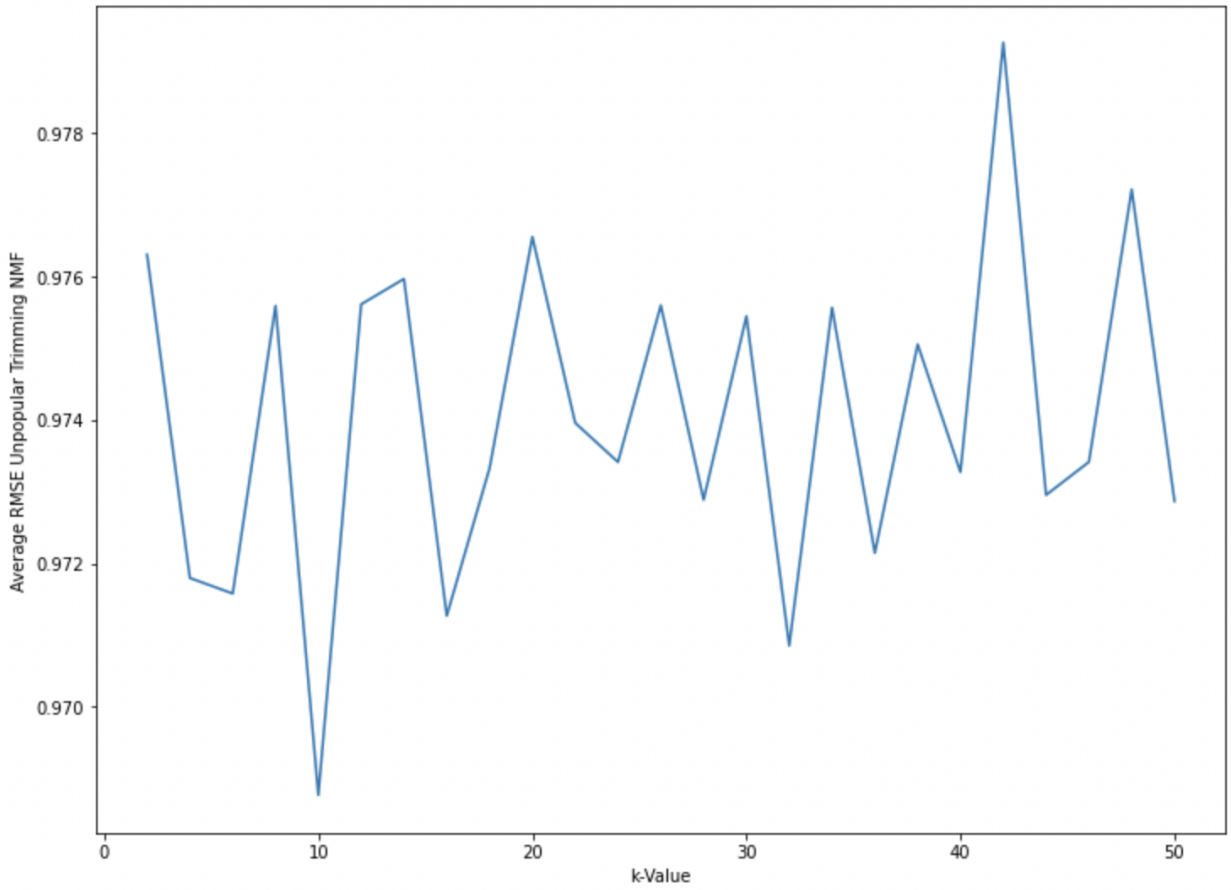


The minimum average RMSE value associated with no trimming was 0.85775

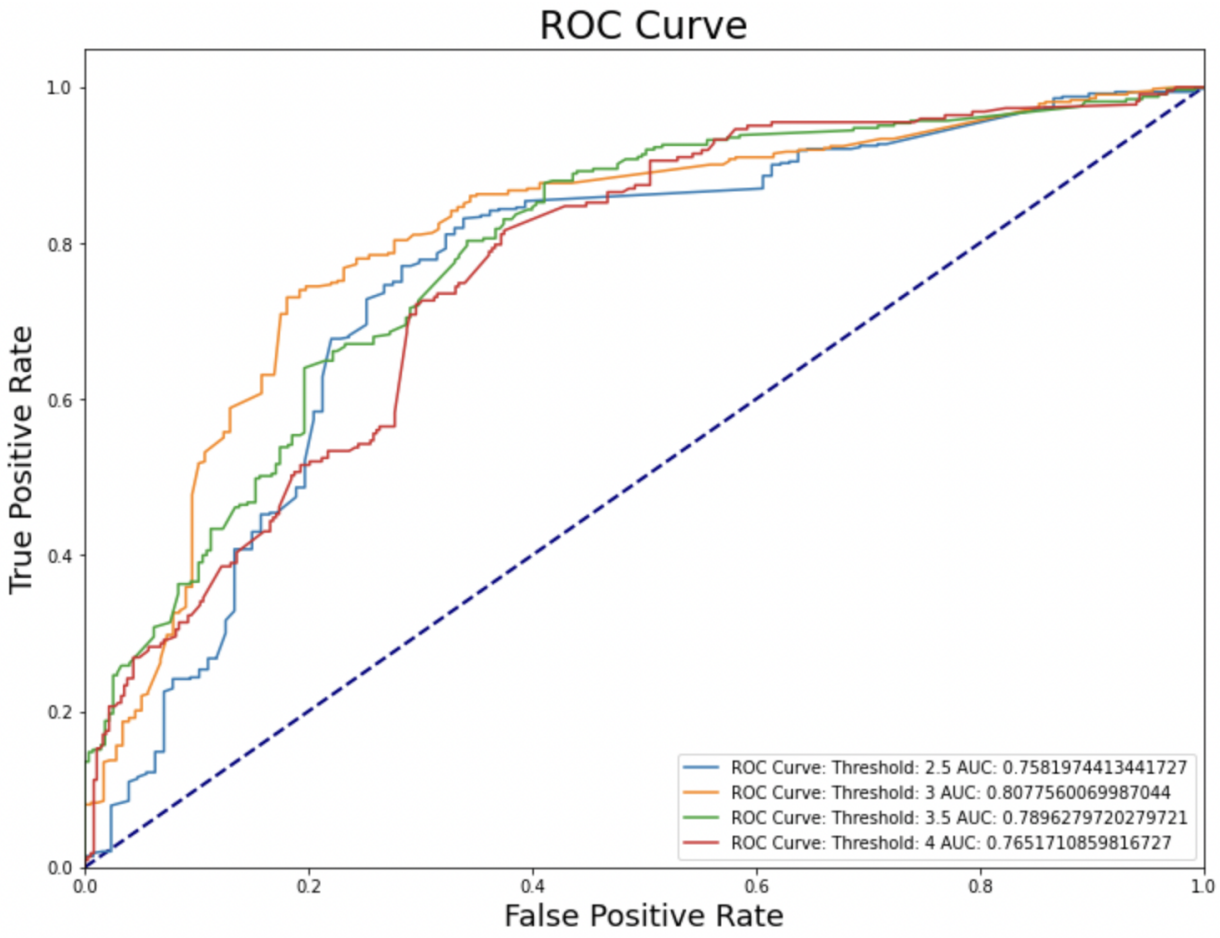


Ratings Threshold	AUC
2.5	0.78962
3.0	0.79244
3.5	0.78740
4.0	0.78114

Unpopular Trimming

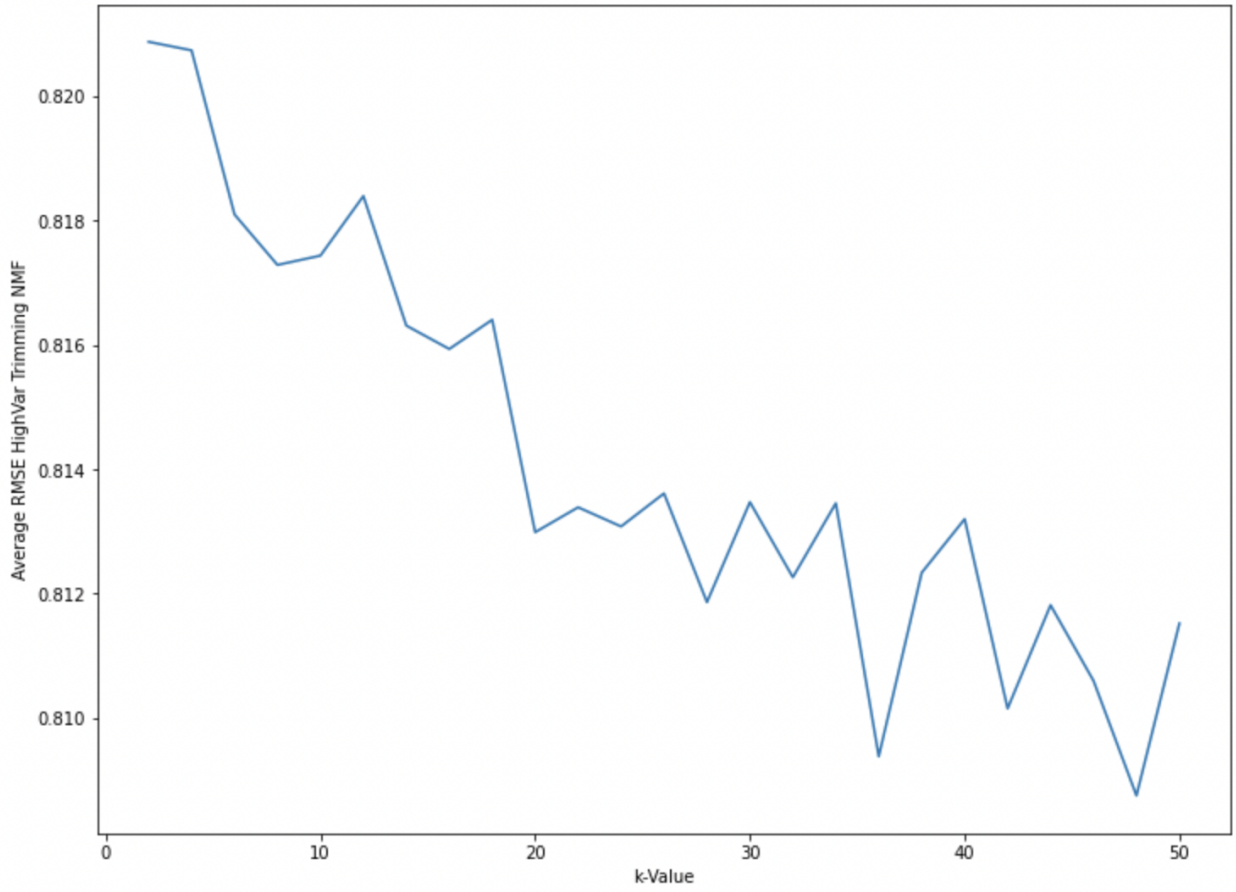


The minimum average RMSE value associated with no trimming was 0.96877

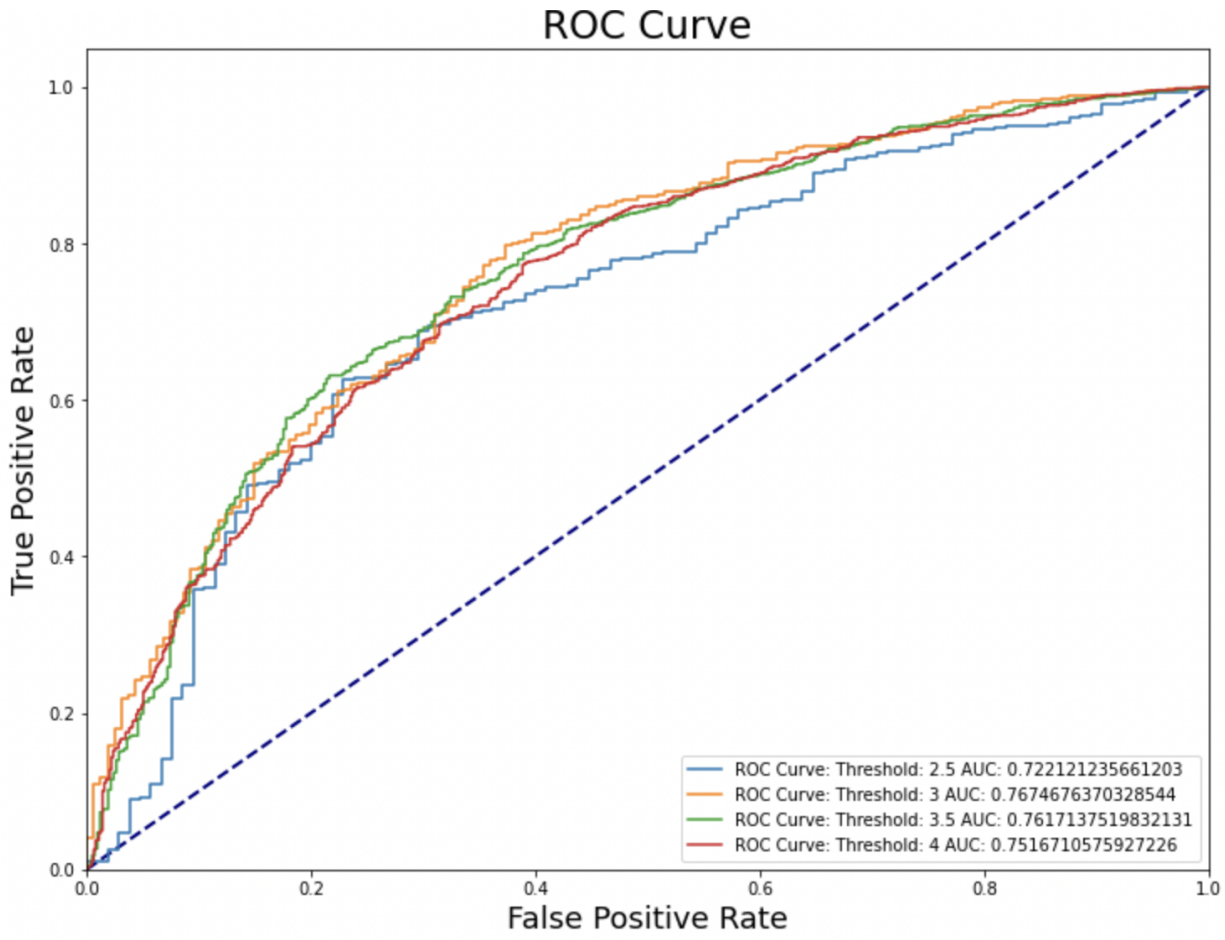


Ratings Threshold	AUC
2.5	0.75820
3.0	0.80776
3.5	0.78963
4.0	0.76517

High Variance Trimming



The minimum average RMSE value associated with no trimming was 0.80875



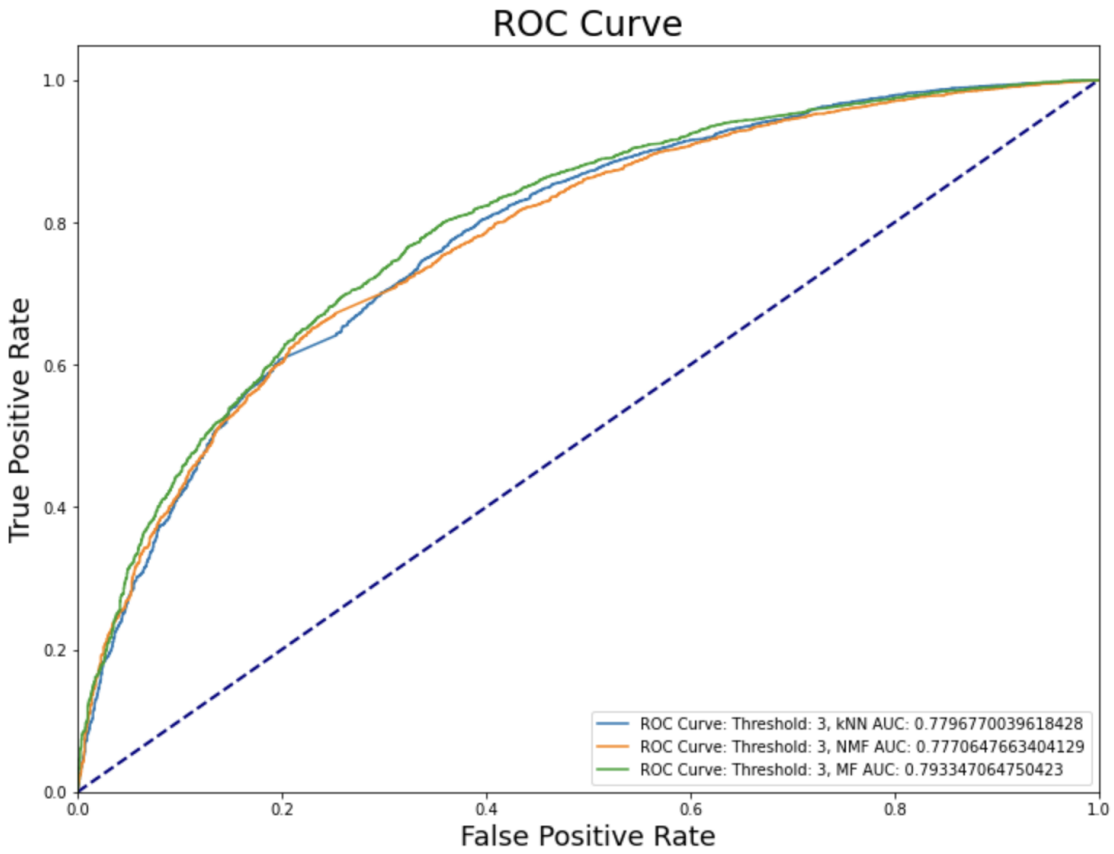
Ratings Threshold	AUC
2.5	0.72212
3.0	0.76747
3.5	0.76174
4.0	0.75167

Question 11:

Naive Collaborative Filter Performance	
Trimming Method	**Average RMSE**
No Trimming	0.93470
Popular Trimming	0.93231
Unpopular Trimming	0.97091
High Variance Trimming	0.91993

The trimming method with the smallest RMSE is the high variance trimming.

Question 12:



The AUC indicates the separability of classes and hence is a direct measure of which classifier performed best. The above figure shows that the model with the highest AUC value was the MF model with bias (0.79335). Second is the kNN method with a value of 0.77968, and lastly NMF with a value of 0.77707.

Question 13:

$S(t)$: The set of items of size t recommended to the user. In this recommended set, ignore (drop) the items for which we don't have a ground truth rating.

G: The set of items liked by the user (ground-truth positives)

Then with the above notation, the expressions for precision and recall are given by equations 8 and 9 respectively

$$Precision(t) = \frac{|S(t) \cap G|}{|S(t)|} \tag{12}$$

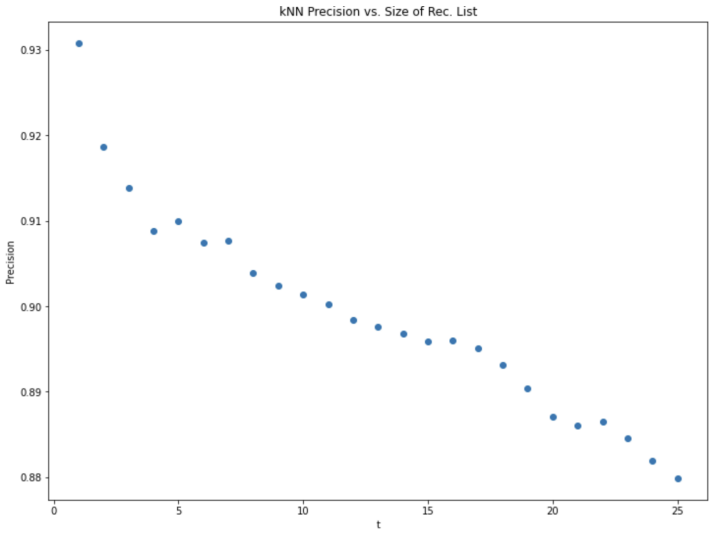
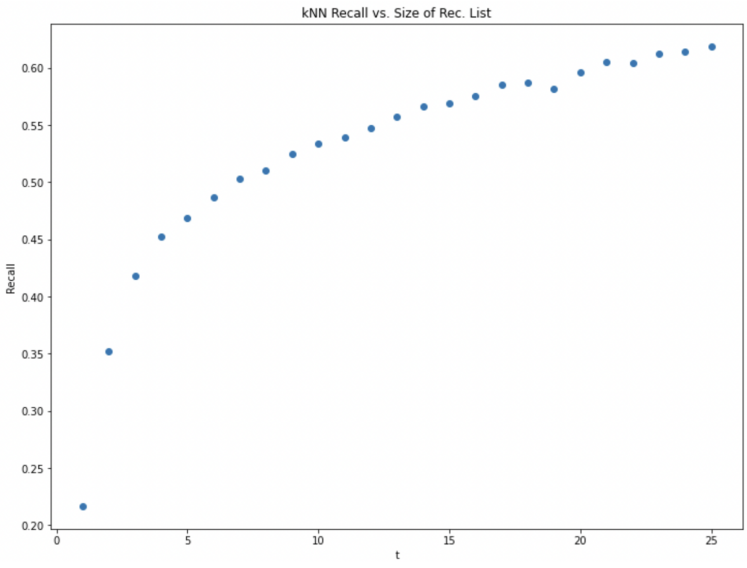
$$Recall(t) = \frac{|S(t) \cap G|}{|G|} \tag{13}$$

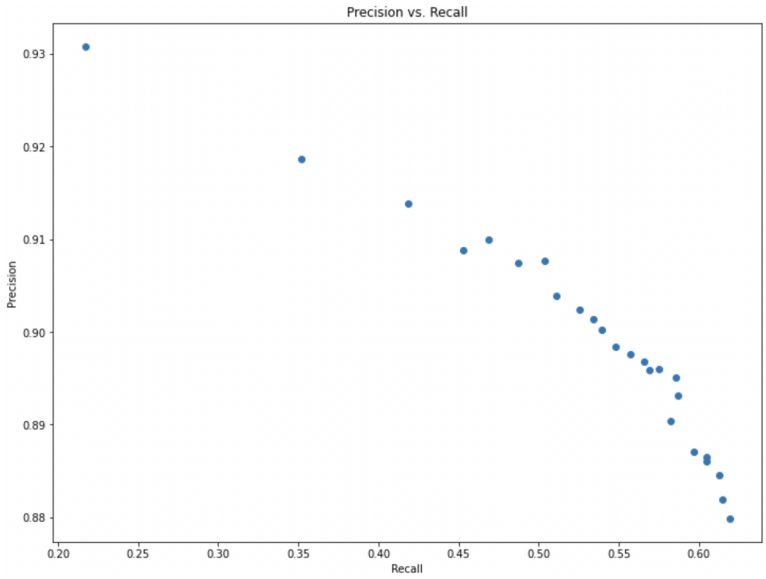
Based on equation (12) above, the precision can be interpreted to mean that of the recommended items to the user, how many of them were actually liked by the user. This is an indication of how accurately the model is performing.

Based on equation (13) above, the recall value is examining the coverage of liked movies by a user by the model. It is taking the intersection of liked movies with the ground truth and reporting it as a proportion of the ground truth.

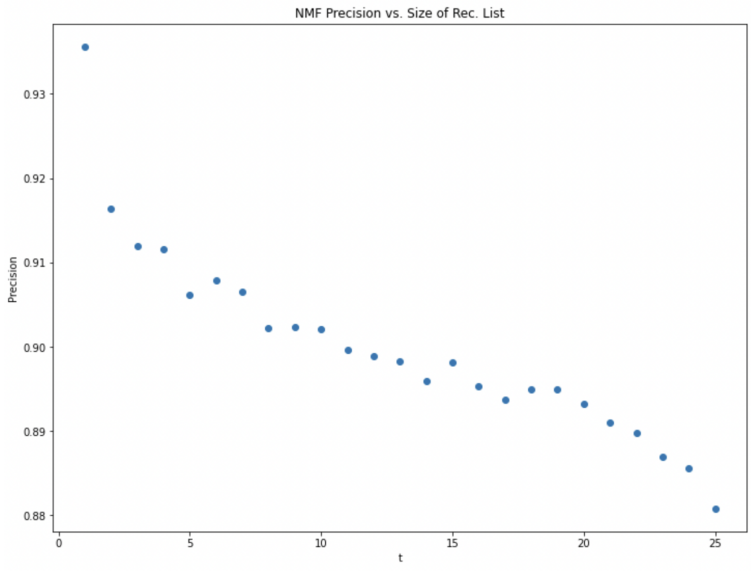
Question 14:

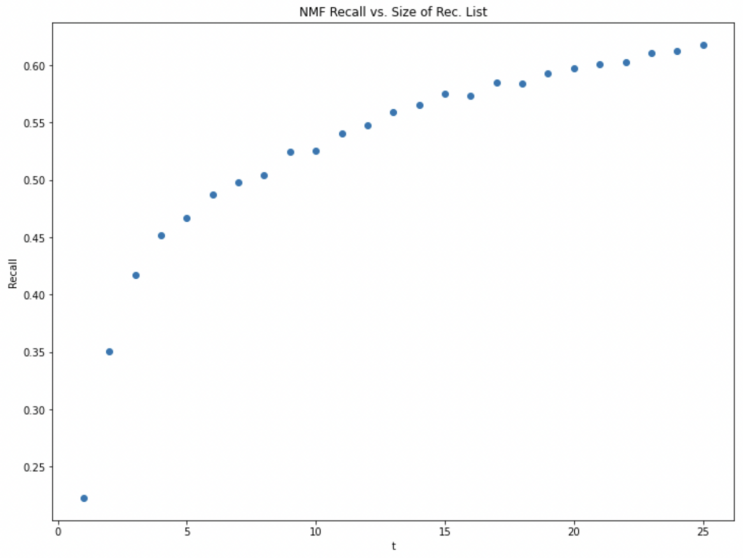
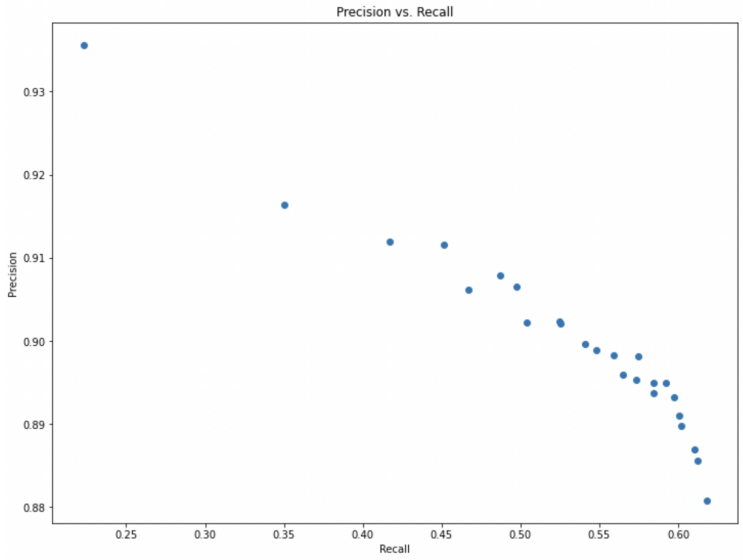
<u>kNN:</u>

Description	Plot
Overall the mean precision appears to be very high with a value of approximately 0.90. As the size of the recommendation list increases, the average precision drops across all folds. The intuition behind this trend might be that as the size of possible movies to choose from increases, the model is more likely to confuse different recommendations.	**kNN Precision vs. t** 
The trend appears to be approaching a recall value of 0.70 as t increases. It becomes harder to improve upon guessing all positive movie recommendations as the available data increases. However, the trend is positive implying that there is still added benefit to adding more information to the model for recall.	**kNN Recall vs. t** 
Based on the interpretations of the last two plots, the trend which appears is to be expected. Recall has a positive correlation	**kNN Precision vs. Recall**

with t whereas Precision has a negative correlation; therefore as Recall increases Precision will decrease.	

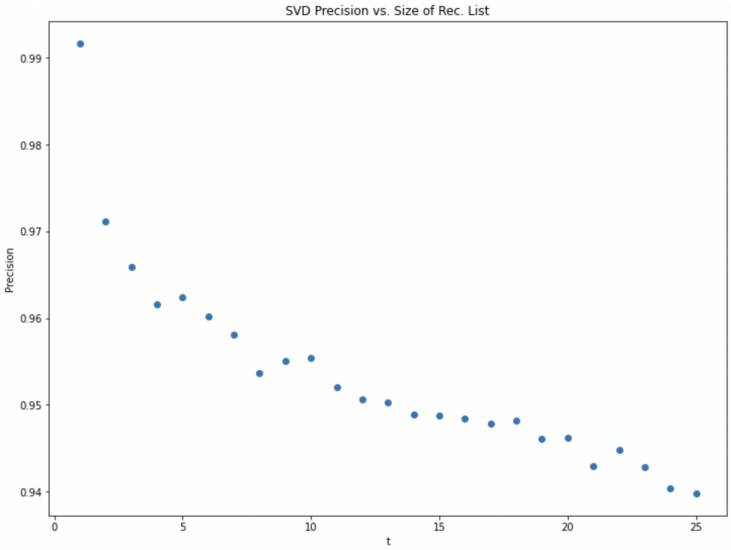
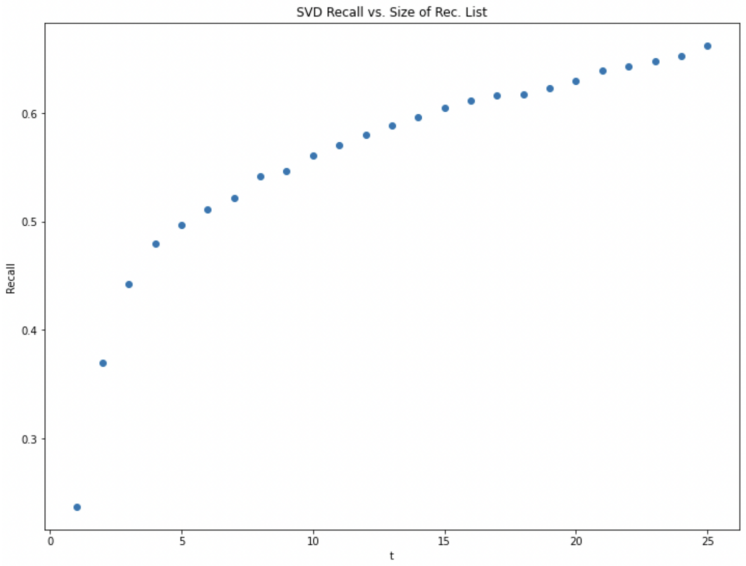
NMF:

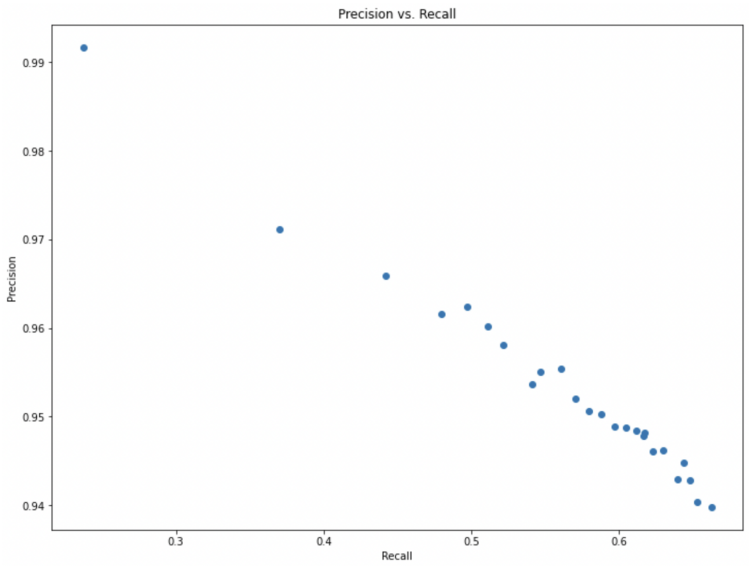
Description	Plot
The trend which appears is very similar to that of kNN. As the size of the recommendation list increases, the average precision drops across all folds. The intuition behind this trend might be that as the size of possible movies to choose from increases, the model is more likely to confuse different recommendations.	**NMF Precision vs. t** 
The trend which appears is very similar to that of kNN. It becomes harder to improve upon guessing all positive movie recommendations as the available data	**NMF Recall vs. t**

increases. However, the trend is positive implying that there is still added benefit to adding more information to the model for recall.	
Based on the interpretations of the last two plots, the trend which appears is to be expected. Recall has a positive correlation with t whereas Precision has a negative correlation; therefore as Recall increases Precision will decrease.	**NMF Precision vs. Recall** 

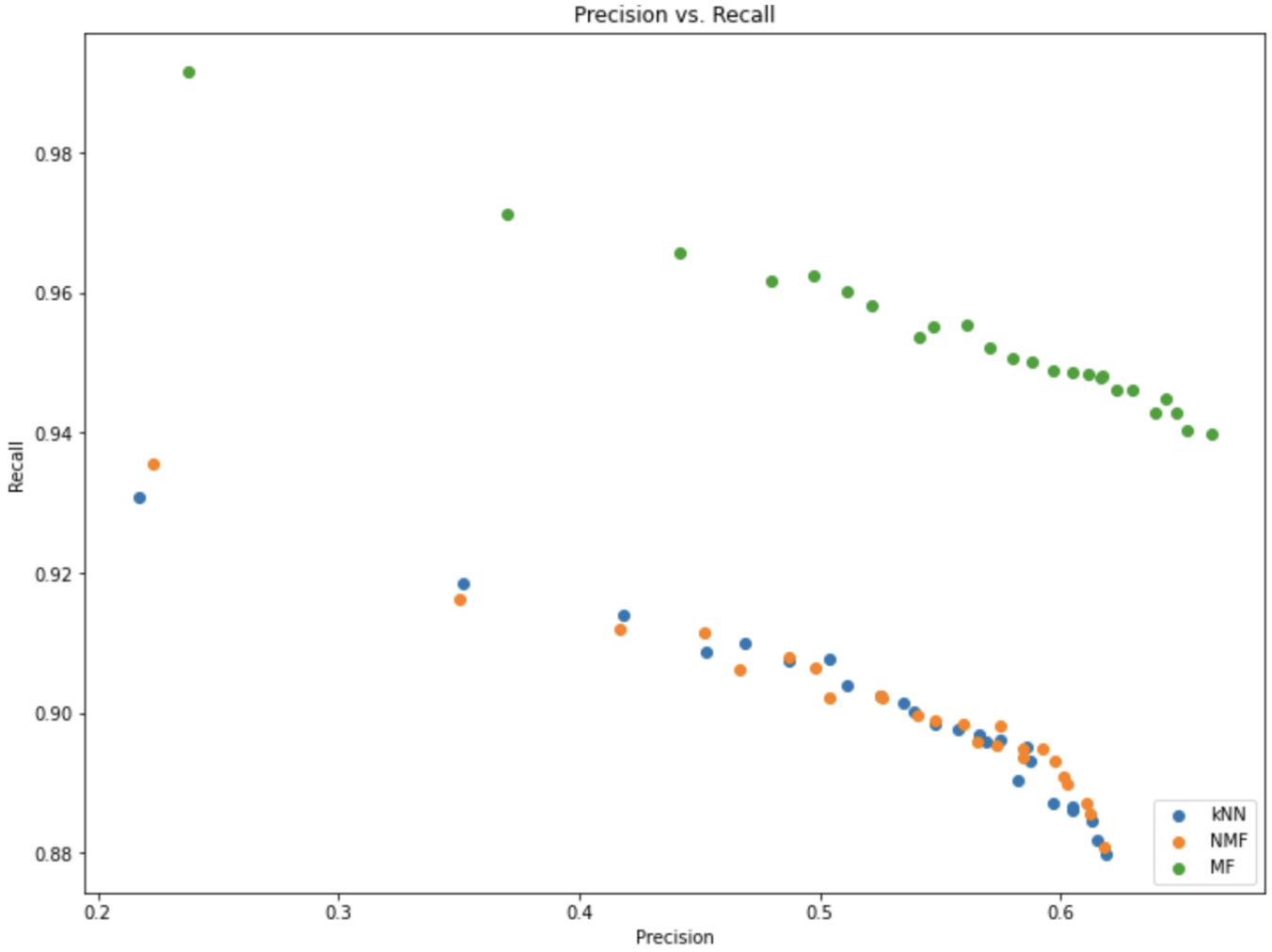
MF:

Description	Plot

The trend which appears is very similar to that of kNN. As the size of the recommendation list increases, the average precision drops across all folds. The intuition behind this trend might be that as the size of possible movies to choose from increases, the model is more likely to confuse different recommendations.	**MF with Bias Precision vs. t** 
The trend which appears is very similar to that of kNN. It becomes harder to improve upon guessing all positive movie recommendations as the available data increases. However, the trend is positive implying that there is still added benefit to adding more information to the model for recall.	**MF with Bias Recall vs. t** 
Based on the interpretations of the last two plots, the trend which appears is to be expected. Recall has a positive correlation with t whereas Precision has a negative correlation; therefore as Recall increases Precision will decrease.	**MF with Bias Precision vs. Recall**


Precision vs. Recall
Precision
Recall
0.99
0.98
0.97
0.96
0.95
0.94
0.3
0.4
0.5
0.6

kNN/NMF/MF with Bias Precision vs Recall



When comparing the plots shown above, MF with bias seems to have a clear advantage over kNN or NMF with respect to recall. This is consistent with the findings shown in earlier sections where MF had better performance.