ECE — Large-Scale Data Mining: Models and Algorithms

Name: Anwesha Chattoraj, Zhaoliang Zheng, Shivam Kumar Panda
Student UID: 405350265, 605432345, 105730045
Due Date: Feb. 28 2022, 11:59 PM **Project:** 2

Question 1

1A

The sparsity of the movie rating dataset is: 0.0169996831

1B

The required histogram showing the frequency of the rating values can be seen in Figure 1. The shape is similar to a normal distribution around 3.5.

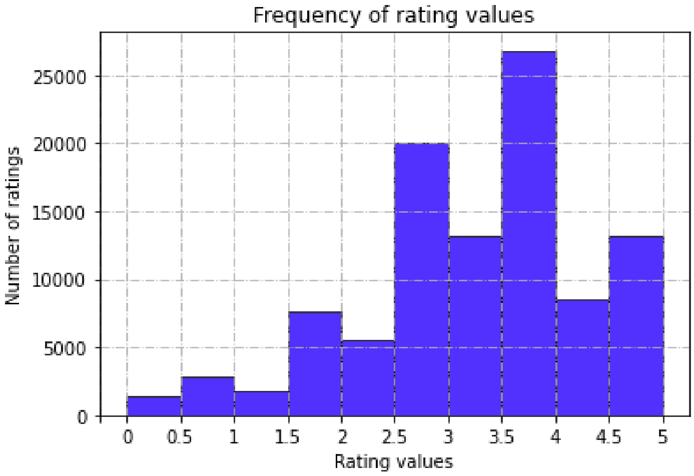


Figure 1: Frequency of rating values with bins of interval 0.5 rating

1C

The required plot for the distribution of the number of ratings received among movies can be seen in the Figure 2

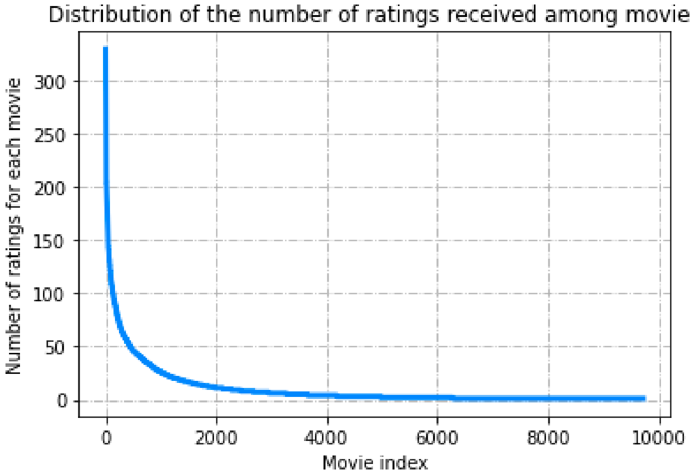


Figure 2: The distribution of the number of ratings received among movies from highest numbers of ratings to lowest

1D

The required plot for the distribution of number of ratings among users can be seen in the Figure 3

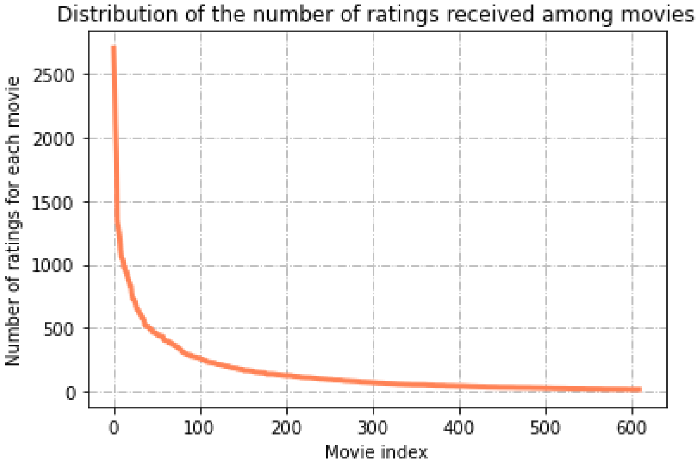


Figure 3: The distribution of the number of ratings given among users from highest numbers of ratings to lowest

1E

There are so many movies but the number of ratings are nothing comparing to that of movies. And there are so many movies that only one or two ratings are received. Even the users with highest number of ratings have rated around 2800 movies out of around 9800 movies. And even the highest number of rated movies are watched by around 350 users out of around 600 users.

1F

The required histogram for the variance of the rating values received by each movie can be seen in the Figure 4. According to this plot we can say that a lot of movies (around 6000 out of 9800) have variance less than 0.5 and even many movies have variance between 0.5 and 1. It should imply that most of the movies are similarly rated by the users. But we need to consider that a lot of movies in this dataset are rated by one or two users or very low number of users (from part 1C). Hence those movies tend to have lower variance compared to the movies which are rated by more number of users.

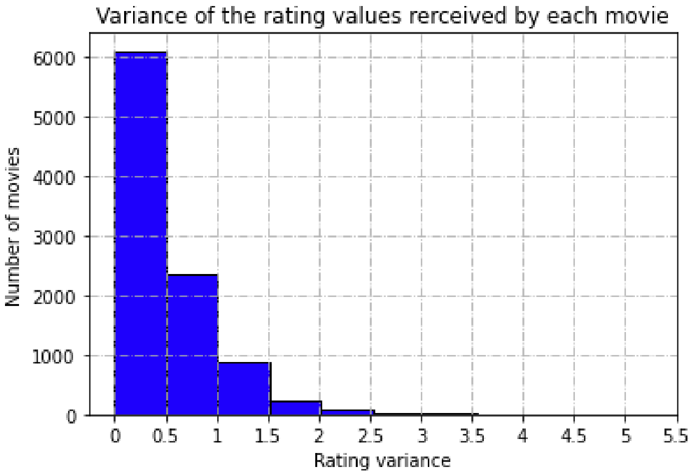


Figure 4: The distribution of the variance of the rating values received by each movie with bins of size 0.5 rating

Question 2

2A

μ_u can be written in terms of I_u and r_{uk} as in the following expression.

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

2B

$I_u \cap I_v$ refers to the set of item indices for which ratings have been specified by both the users (u and v). Yes it is possible that $I_u \cap I_v = \emptyset$. It means that there are no such common items for which both user u and user v have given their ratings.

Question 3

A user's opinion and feedback about a movie is subjective. Further the ratings 0 to 5 may mean different for different users. For example person A can give rating 4 for a ”good movie” and 4.5 for an ”excellent movie” according to him. Whereas another person B can give rating 3.5 for a ”good movie” and 4 for an ”excellent movie” according to him. This means person B is more critical compared to person A. Hence users may have different bias when they convert their subjective ”feelings” and ”opinions” of any movie into objective ”ratings”. Hence mean centring is important to normalize this bias among the users.

Question 4

The required plot of average RMSE (Y-axis) against k (X-axis) and average MAE (Y-axis) against k (X-axis) can be seen in the Figure 5

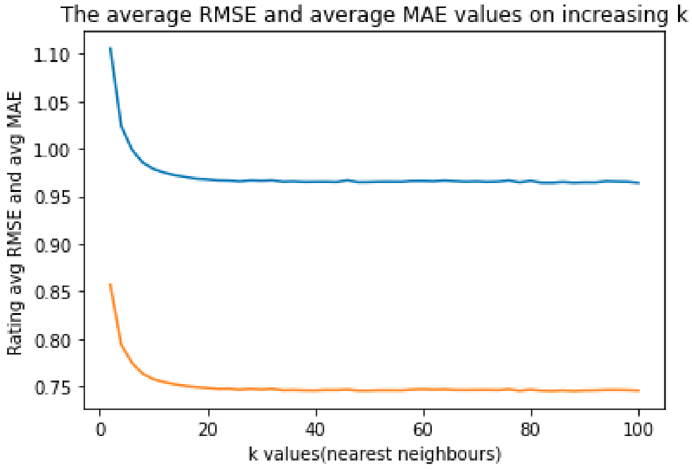


Figure 5: The average RMSE and average MAE vs increasing value of k (nearest neighbours)

Question 5

With converging error of less than 0.0005 we get $minimum\_k = 22$. The steady state value of average RMSE is around 0.9642354 and the steady state value of average MAE is around 0.7453425

Question 6

6A: Popular Movie Trimming

The required plot of average RMSE (Y-axis) against k (X-axis) can be seen in the Figure 6. The minimum average value of RMSE is: 0.9213966405429034.

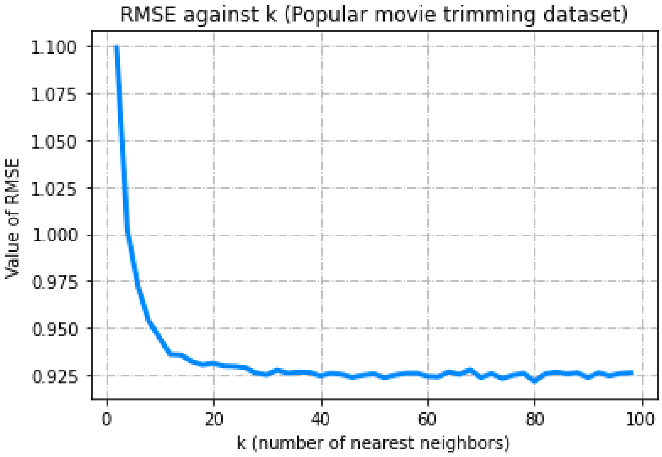
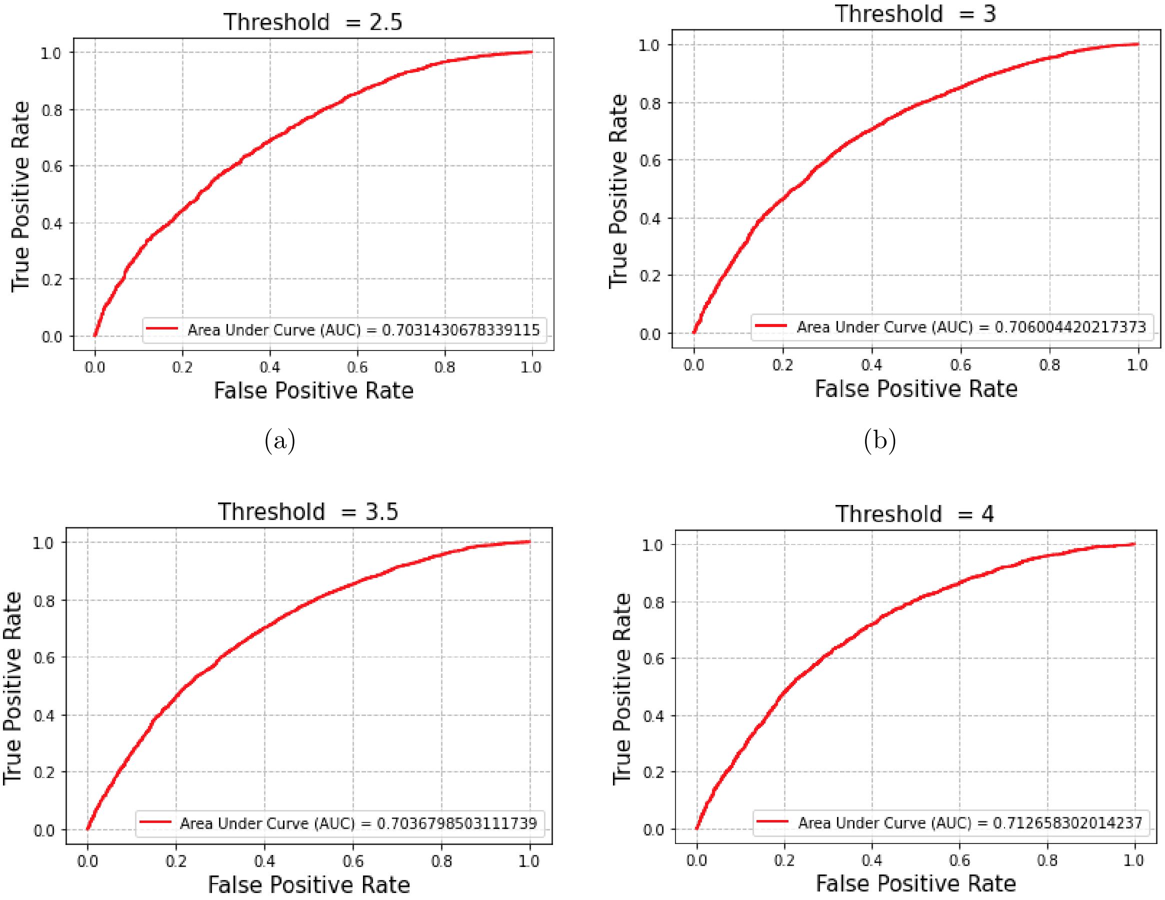


Figure 6: The average RMSE vs increasing values of k (nearest neighbours) for popular movie trimming dataset

The required ROC curves for the k-NN collaborative filters for threshold values $[2.5, 3, 3.5, 4]$ can be seen in the Figure 7



(a) (b)

(c) (d)

Figure 7: The ROC curves with threshold values (a) 2.5 (b) 3 (c) 3.5 (d) 4

6B: Unopular Movie Trimming

The required plot of average RMSE (Y-axis) against k (X-axis) can be seen in the Figure 8. The minimum average value of RMSE is: 1.0406948656358666.

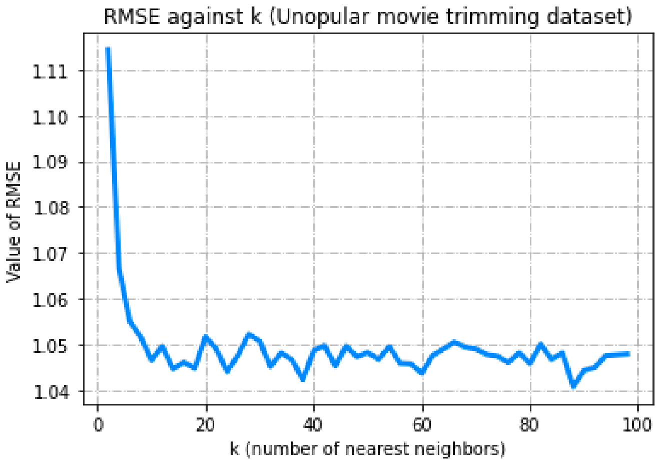
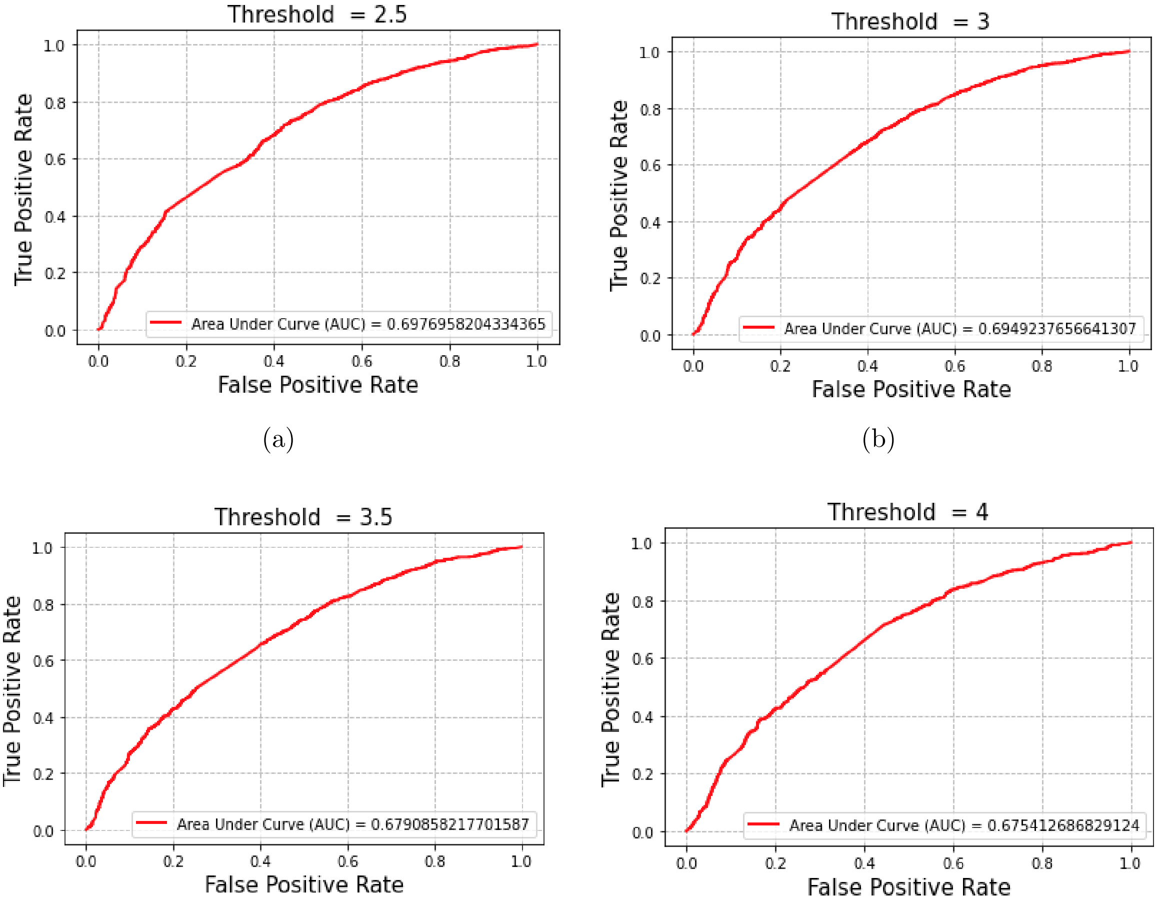


Figure 8: The average RMSE vs increasing values of k (nearest neighbours) for unpopular movie trimming dataset

The required ROC curves for the k-NN collaborative filters for threshold values $[2.5, 3, 3.5, 4]$ can be seen in the Figure 9



(a)

(b)

(c)

(d)

Figure 9: The ROC curves with threshold values (a) 2.5 (b) 3 (c) 3.5 (d) 4

6C: High variance Movie Trimming

The required plot of average RMSE (Y-axis) against k (X-axis) can be seen in the Figure 10. The minimum average value of RMSE is: 1.53843749095183.

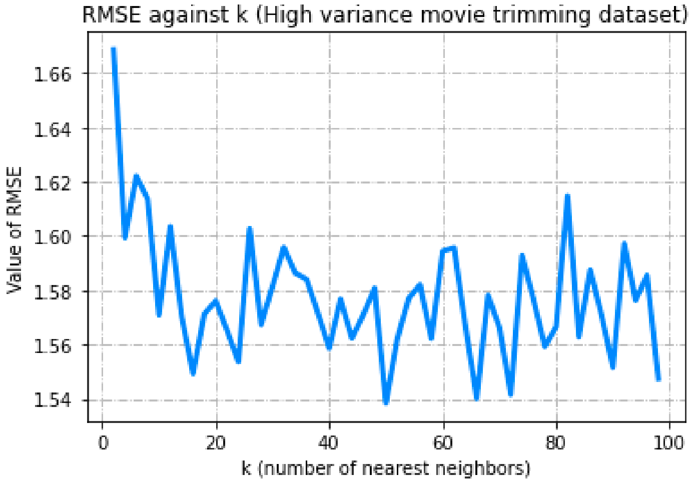
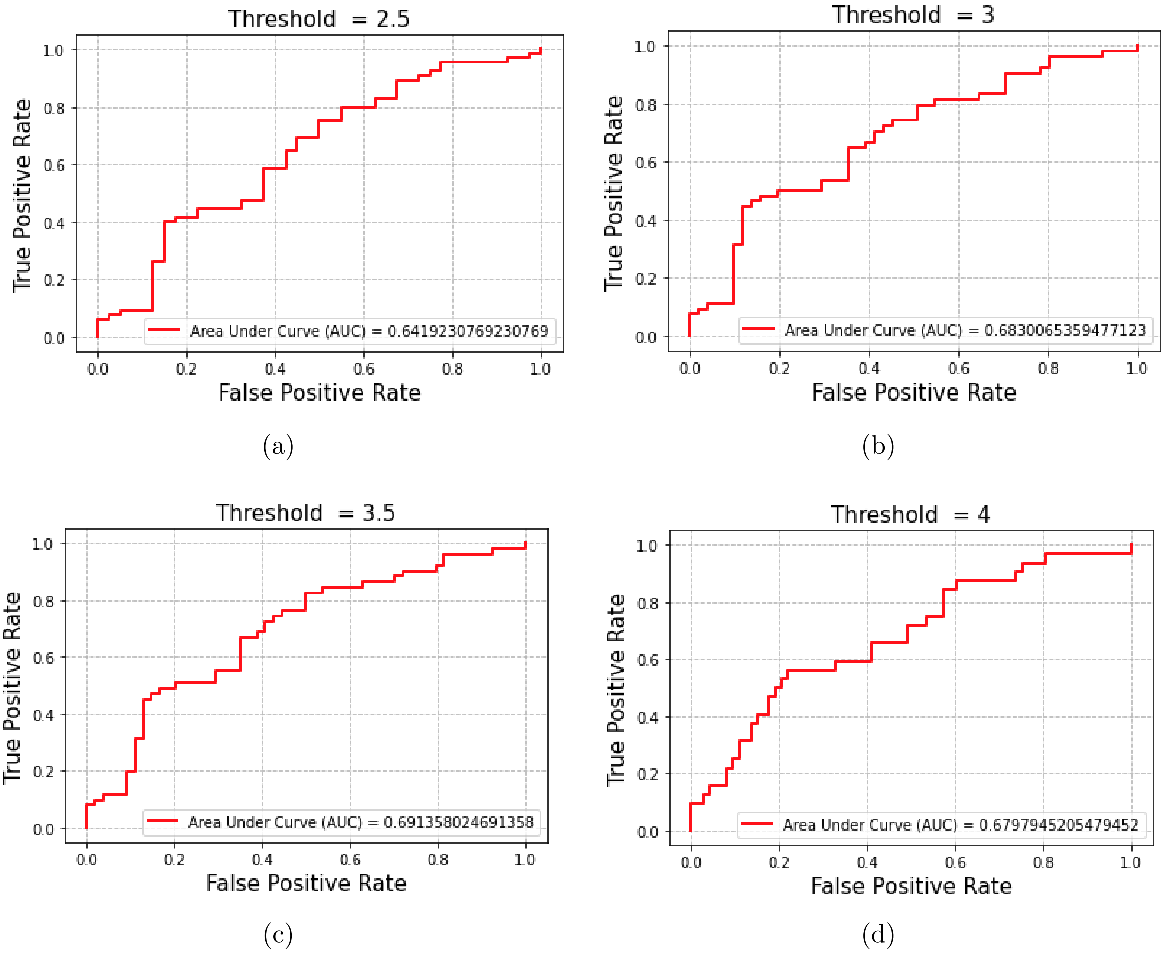


Figure 10: The average RMSE vs increasing values of k (nearest neighbours) for high variance movie trimming dataset

The required ROC curves for the k-NN collaborative filters for threshold values $[2.5, 3, 3.5, 4]$ can be seen in the Figure 11



(a) (b) (c) (d)

Figure 11: The ROC curves with threshold values (a) 2.5 (b) 3 (c) 3.5 (d) 4

Question 7

$$\begin{array}{ll} \underset{U,V}{\text{minimize}} & \sum_{i=1}^{m}\sum_{j=1}^{n} W_{ij}\left(r_{ij} - \left(UV^T\right)_{ij}\right)^2 + \lambda\|U\|_F^2 + \lambda\|V\|_F^2 \\ \text{subject to} & U \geq 0, V \geq 0 \end{array} \tag{1}$$

For U fixed, formulate it as a least-squares problem.

To check convexity:

Considering the scalar case where $m = n = 1$. Writing r, U, V scalars.

$$\min_{u,w\geq 0}(r-uv)^2 = \min_{u,v\geq 0} r^2 - 2ruv + u^2v^2 \tag{2}$$

The Gradient of this is :

$$\nabla\phi_x(u,v) = \begin{bmatrix} 2uv^2 - 2rv \\ 2u^2v - 2ru \end{bmatrix} \tag{3}$$

And the Hessian

$$\nabla^2\phi_x(u,v) = \begin{bmatrix} 2v^2 & 4uv - 2r \\ 4uv - 2r & 2u^2 \end{bmatrix} \tag{4}$$

To prove convexity, the Hessian must be positive Semi definite. The determinant must be > 0.

$$-12u^2v^2 > 4r^2 + 8urv \tag{5}$$

And this hold true iff

$$r > 2vu \tag{6}$$

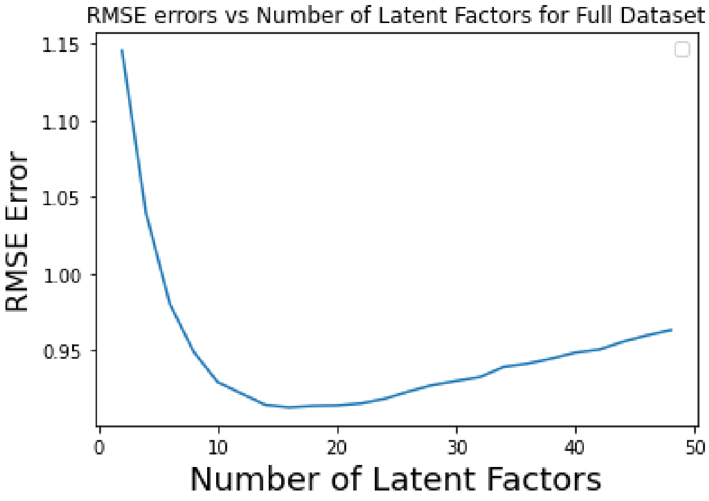
Hence it does not hold true for all values of r, u and v, hence this is not a Convex optimization problem Fixing U reduces it to:

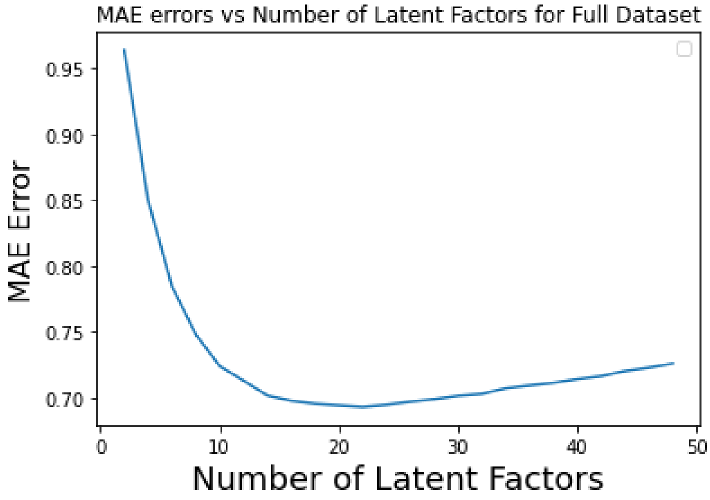
$$\min_V \sum_{i=1}^{m} \sum_{j=1}^{n} W_{ij} \left(r_{ij} - u_i^T v_j \right)^2 \tag{7}$$

Which as we can see is convex, and is in fact a least squares formulation

Question 8

8A



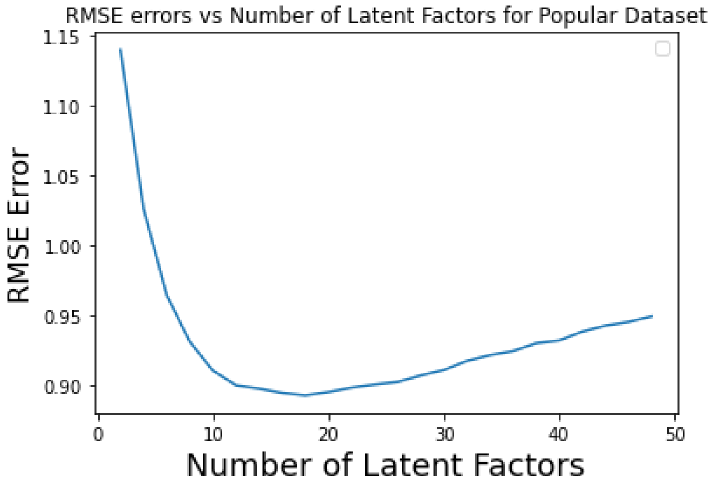


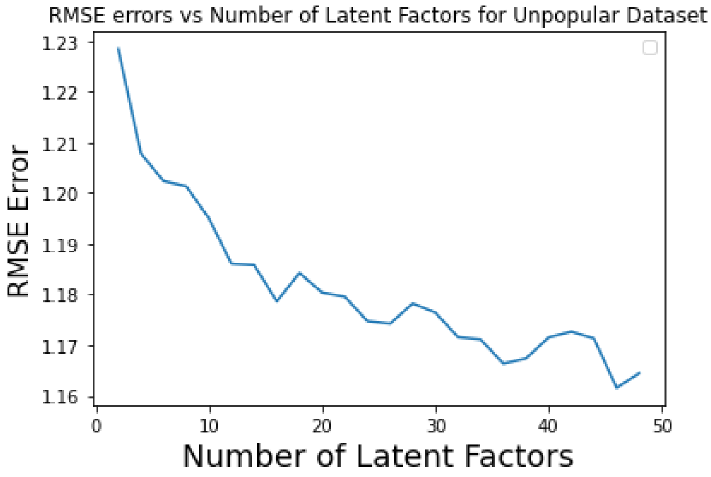
8B

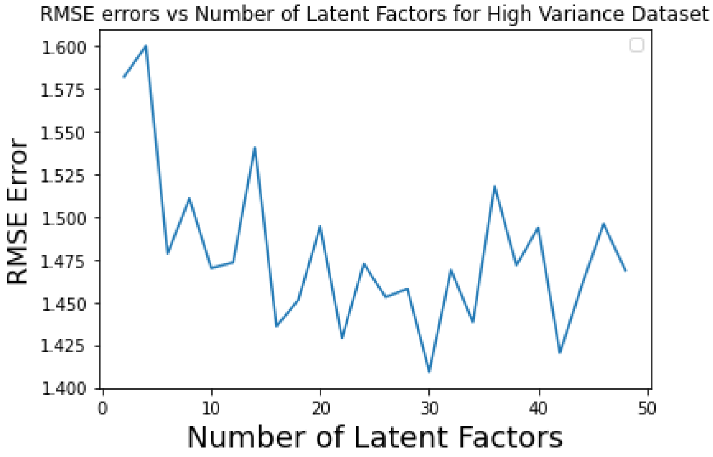
- Optimal number of latent factors for NMF via RMSE Error is 16 and MAE error is 22
- No, the number of movie genres if 951 but the optimal latent factors are 16 or 22 here

8C

Performance on trimmed test set subsets







Optimal number of latent factors for NMF on Popular Test set 18 Unpopular Test Set 46 High Variance Test Set 30

	Popular	Unpopular	High-Variance
Latent Factors	18	46	30

8D

Plotting the ROC curves for the NMF-based collaborative filter designed in part A for threshold values [2.5, 3, 3.5, 4]

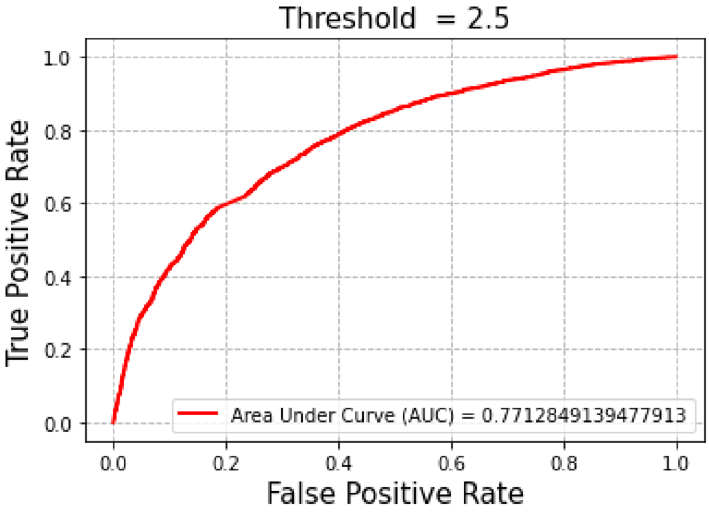


Figure 12: Optimal number of latent factors found via RMSE:16

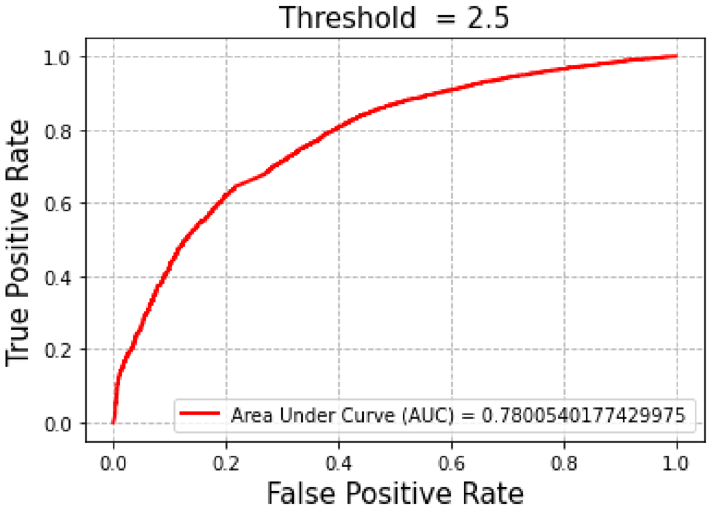


Figure 13: Optimal number of latent factors found via MAE:22

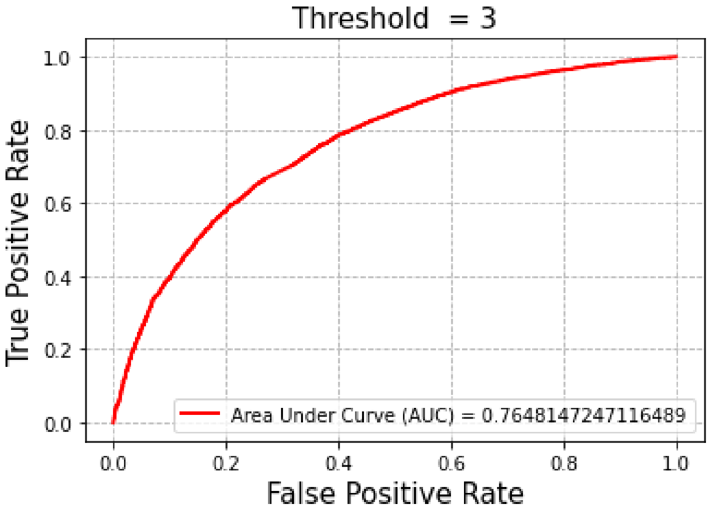


Figure 14: Optimal number of latent factors found via RMSE:16

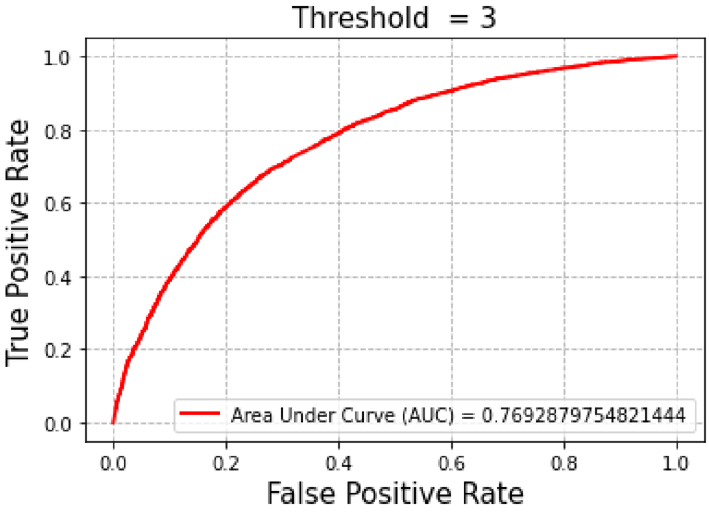


Figure 15: Optimal number of latent factors found via MAE:22

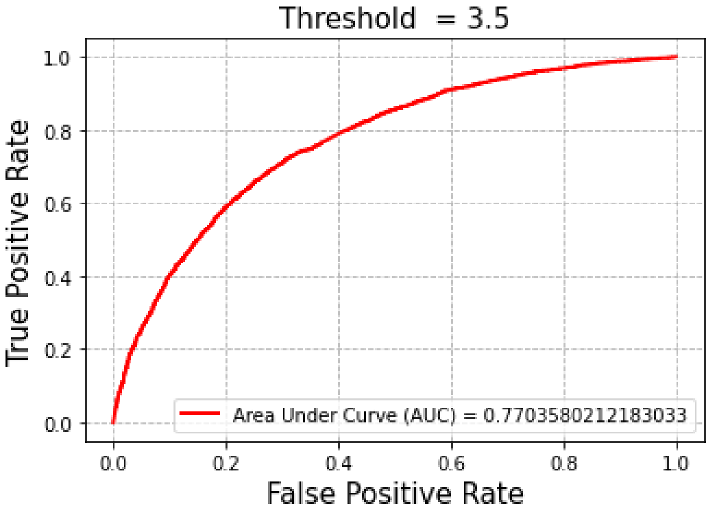


Figure 16: Optimal number of latent factors found via MAE:22

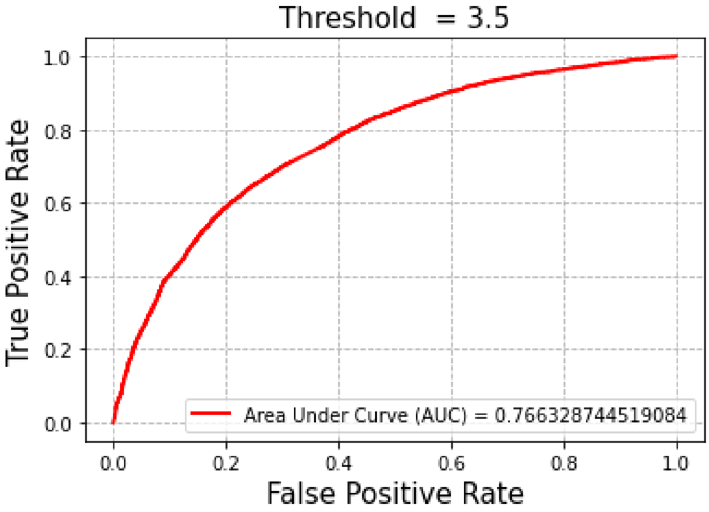


Figure 17: Optimal number of latent factors found via RMSE:16

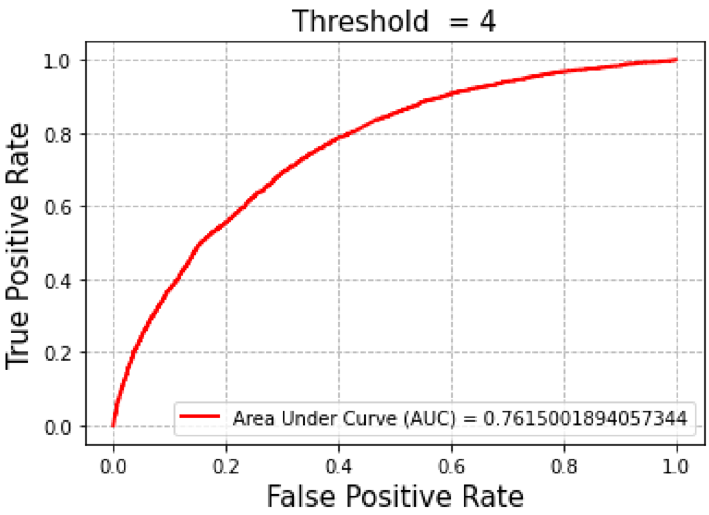


Figure 18: Optimal number of latent factors found via RMSE:16

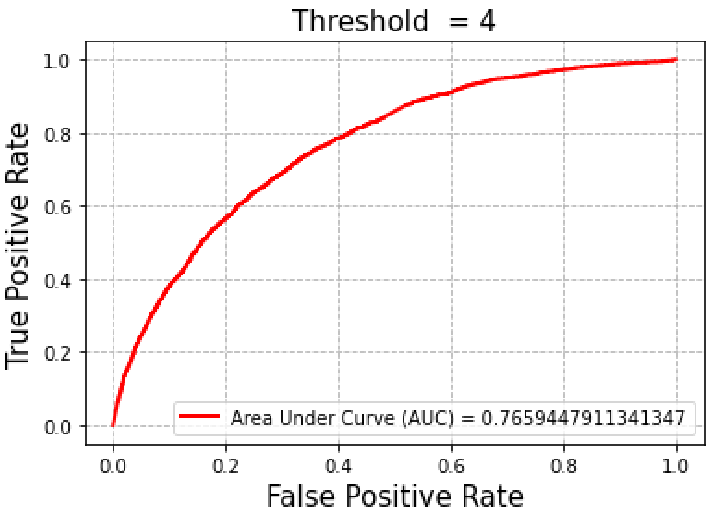
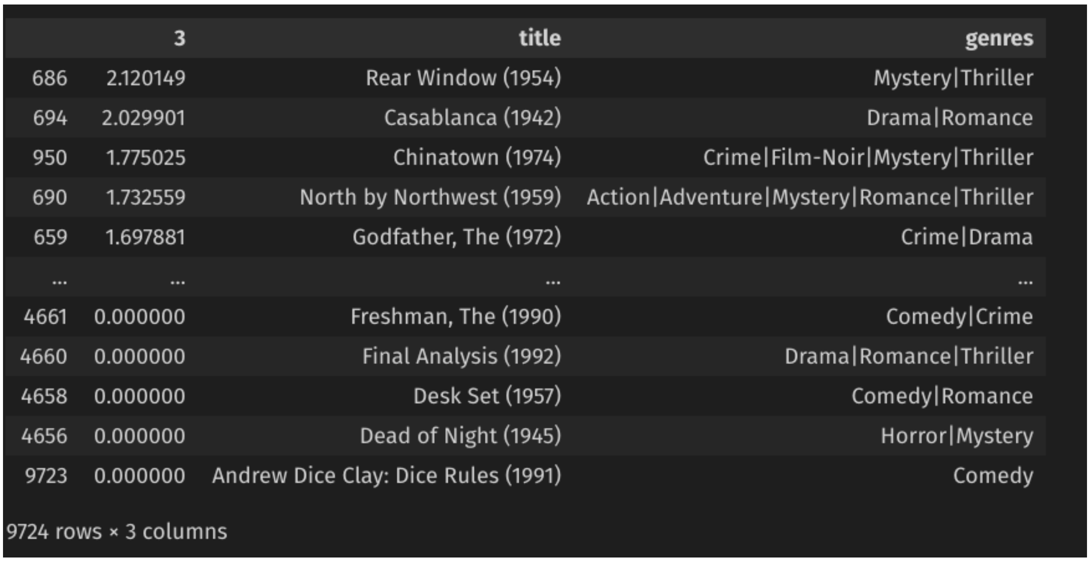


Figure 19: Optimal number of latent factors found via MAE:22

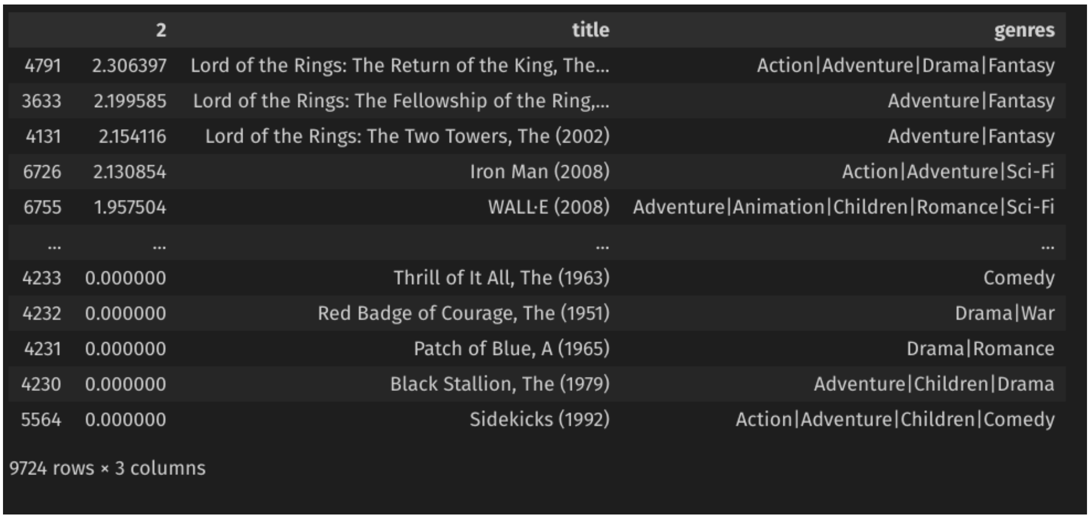
Question 9

Interpreting the NMF model.
Here we have evaluated the latent factors that arise from matrix factorization. We are looking at the Moive-latent factor interations for $k = 20$ (20 latent factors). Then we have sorted each of the latent factors in descending order to see how the latent factors have grouped together information about the movies.



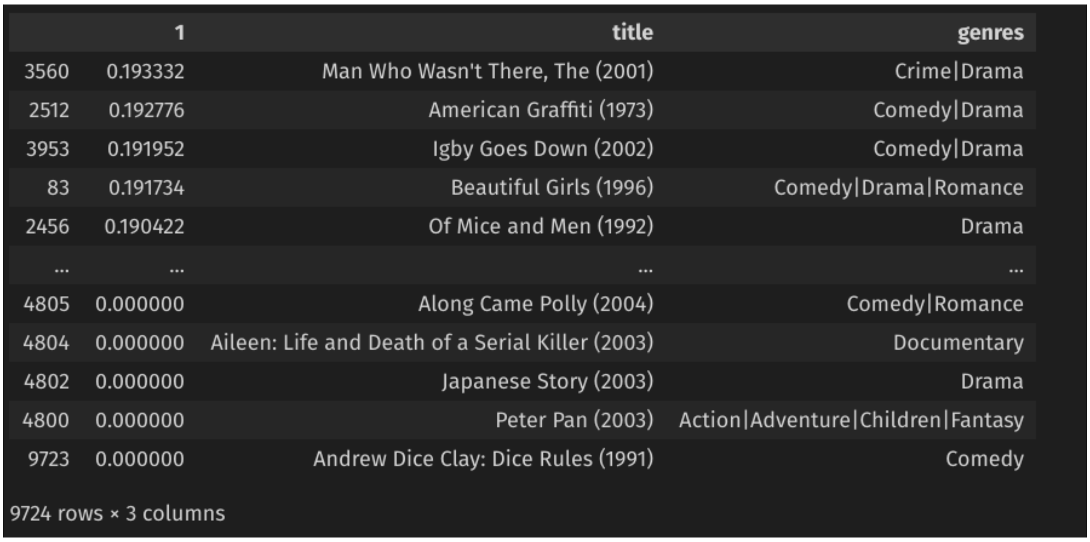
	3	title	genres
686	2.120149	Rear Window (1954)	Mystery|Thriller
694	2.029901	Casablanca (1942)	Drama|Romance
950	1.775025	Chinatown (1974)	Crime|Film-Noir|Mystery|Thriller
690	1.732559	North by Northwest (1959)	Action|Adventure|Mystery|Romance|Thriller
659	1.697881	Godfather, The (1972)	Crime|Drama
...	...	...	...
4661	0.000000	Freshman, The (1990)	Comedy|Crime
4660	0.000000	Final Analysis (1992)	Drama|Romance|Thriller
4658	0.000000	Desk Set (1957)	Comedy|Romance
4656	0.000000	Dead of Night (1945)	Horror|Mystery
9723	0.000000	Andrew Dice Clay: Dice Rules (1991)	Comedy

9724 rows × 3 columns



	2	title	genres
4791	2.306397	Lord of the Rings: The Return of the King, The...	Action|Adventure|Drama|Fantasy
3633	2.199585	Lord of the Rings: The Fellowship of the Ring,...	Adventure|Fantasy
4131	2.154116	Lord of the Rings: The Two Towers, The (2002)	Adventure|Fantasy
6726	2.130854	Iron Man (2008)	Action|Adventure|Sci-Fi
6755	1.957504	WALL·E (2008)	Adventure|Animation|Children|Romance|Sci-Fi
...	...	...	...
4233	0.000000	Thrill of It All, The (1963)	Comedy
4232	0.000000	Red Badge of Courage, The (1951)	Drama|War
4231	0.000000	Patch of Blue, A (1965)	Drama|Romance
4230	0.000000	Black Stallion, The (1979)	Adventure|Children|Drama
5564	0.000000	Sidekicks (1992)	Action|Adventure|Children|Comedy

9724 rows × 3 columns



	1	title	genres
3560	0.193332	Man Who Wasn't There, The (2001)	Crime|Drama
2512	0.192776	American Graffiti (1973)	Comedy|Drama
3953	0.191952	Igby Goes Down (2002)	Comedy|Drama
83	0.191734	Beautiful Girls (1996)	Comedy|Drama|Romance
2456	0.190422	Of Mice and Men (1992)	Drama
...	...	...	...
4805	0.000000	Along Came Polly (2004)	Comedy|Romance
4804	0.000000	Aileen: Life and Death of a Serial Killer (2003)	Documentary
4802	0.000000	Japanese Story (2003)	Drama
4800	0.000000	Peter Pan (2003)	Action|Adventure|Children|Fantasy
9723	0.000000	Andrew Dice Clay: Dice Rules (1991)	Comedy

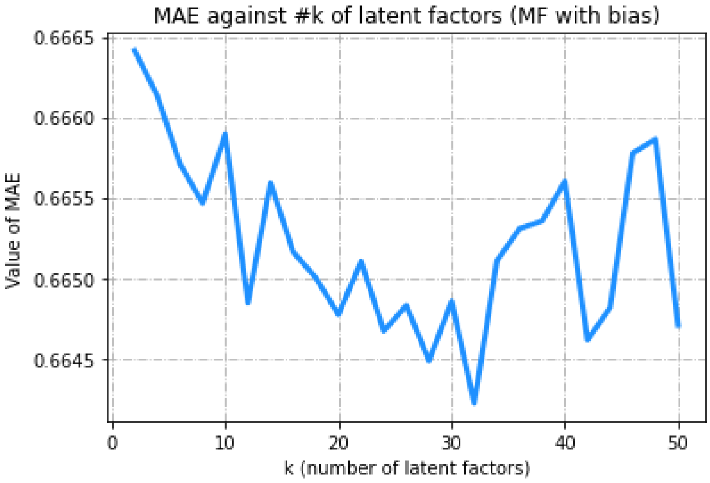
9724 rows × 3 columns

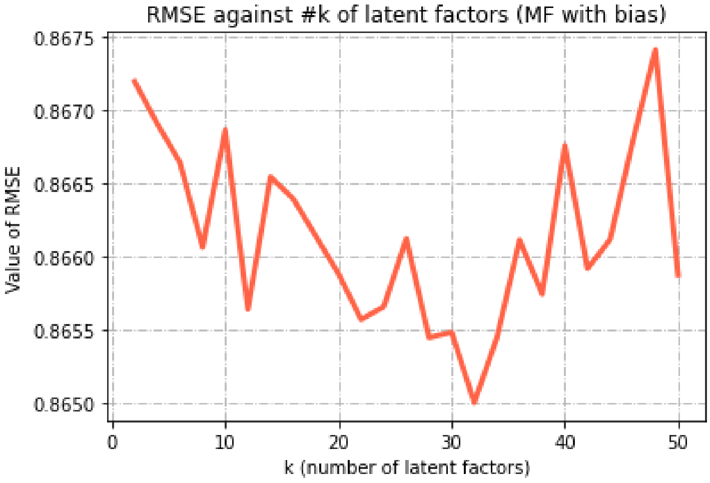
These are some the sorted columns of the movie-latent factors matrices. As you can clearly see, the latent factors seem to have grouped together certain genres/groups of genres with similar characteristics. E.g: Comedy, Drama, Romance are grouped together, Adventure, Fantasy, SciFi are occurring together etc

Question 10

10A

The plots of average MAE and RMSE against k are shown as below:





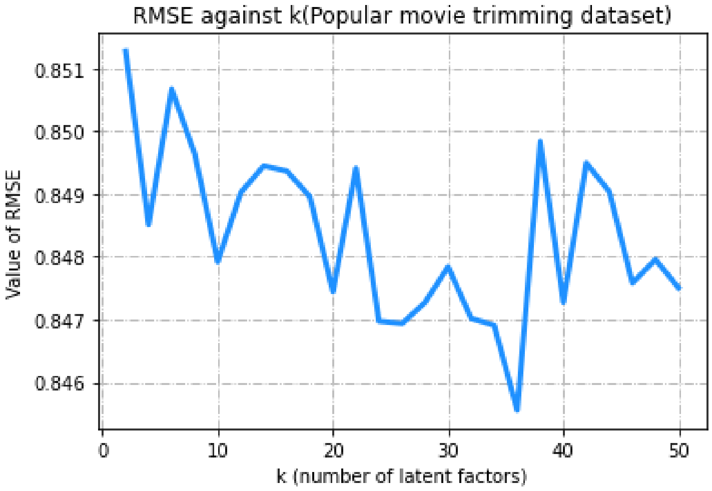
10B

Based on the plots in question B, the optimal number of latent factors are shown in the following table:

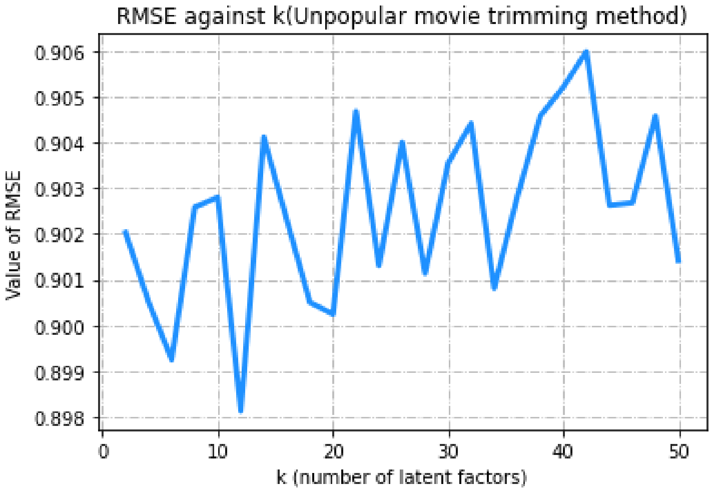
	MAE	RMSE
OPTIMAL k	32	32
Minimum Average	0.6642197646627767	0.8649991036150022

The genres of the movies is 20, which is not same as the optimal number of latent factor k: 32.

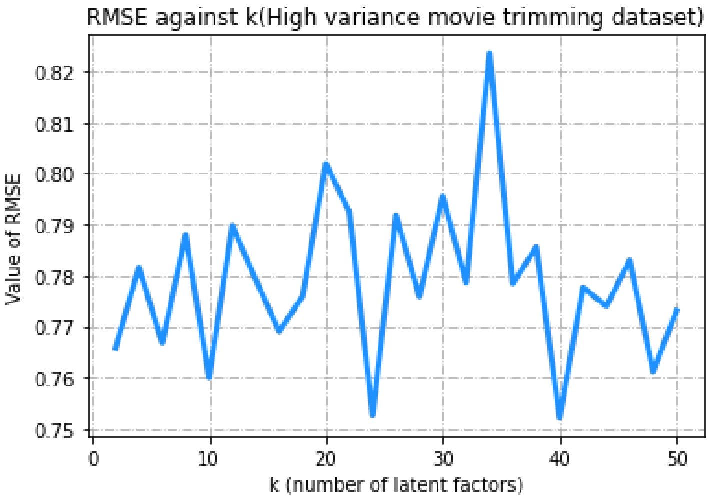
10C



The minimum average RMSE for Popular trimming dataset: 0.8455542687780675



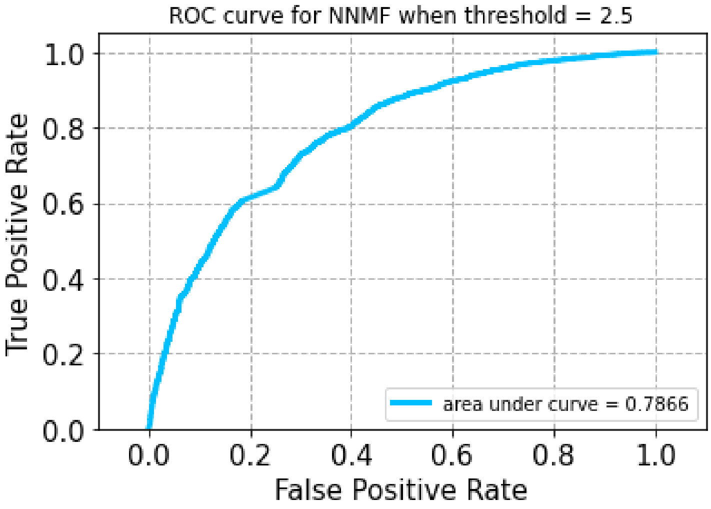
The minimum average RMSE for Unpopular trimming dataset: 0.898120483149996

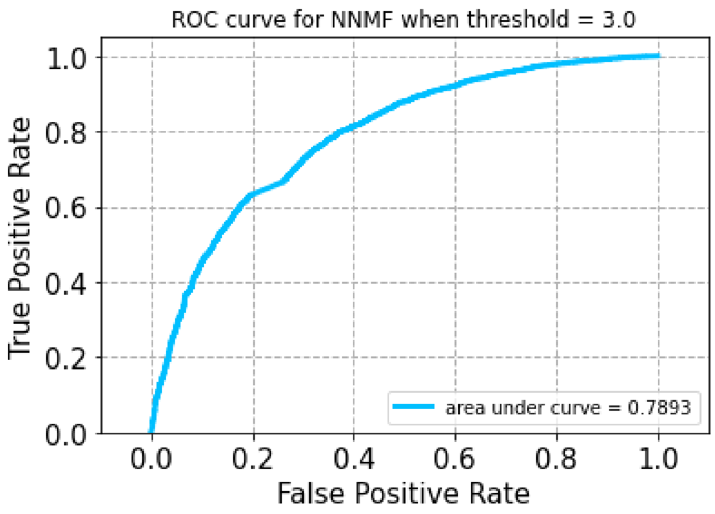


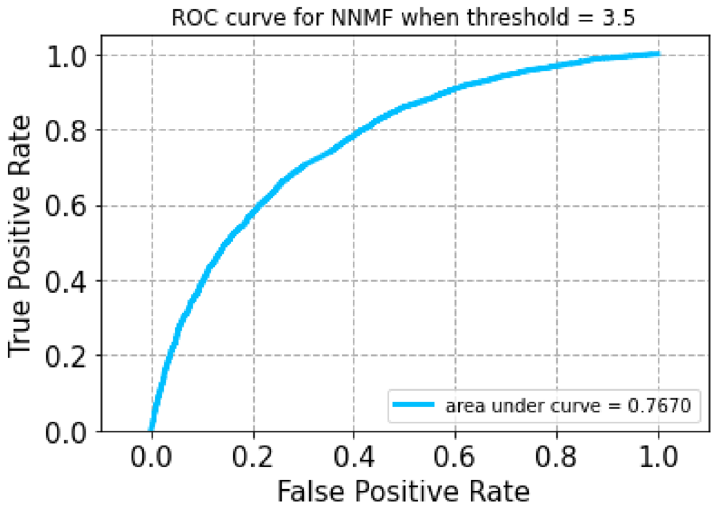
The minimum average RMSE for high variance trimming dataset: 0.7521811436833071

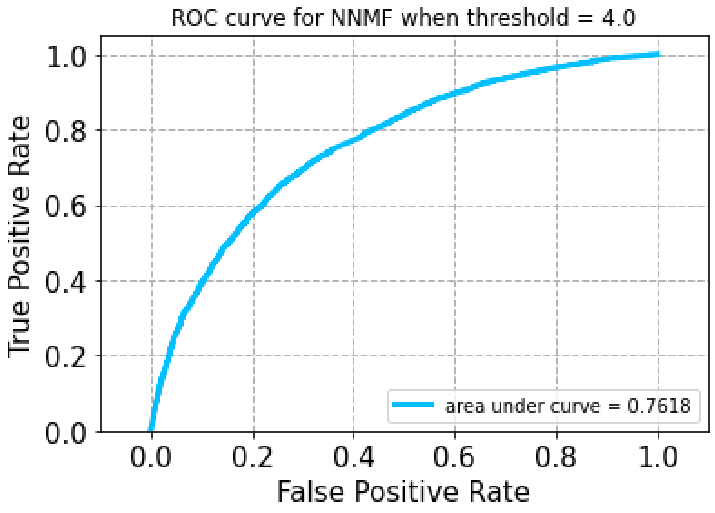
10D

For each of the plot, the area under the curve was reported in the legend of the plot as below:




ROC curve for NNMF when threshold = 3.0
True Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
area under curve = 0.7893


ROC curve for NNMF when threshold = 3.5
True Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
area under curve = 0.7670


ROC curve for NNMF when threshold = 4.0
True Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
area under curve = 0.7618

Question 11

Average RMSE

The average RMSE result is:

$$Avg(RMSE) = 0.9412426150480979$$

Performance on Test set subsets

We designed a naive collaborative filter to predict the ratings of movies for each of Popular, Unpopular and High-Variance Test subsets. The results are shown in the following table:

	Popular	Unpopular	High-Variance
RMSE	0.9312063941499369	0.9634131366983443	0.8668596555993855

Question 12

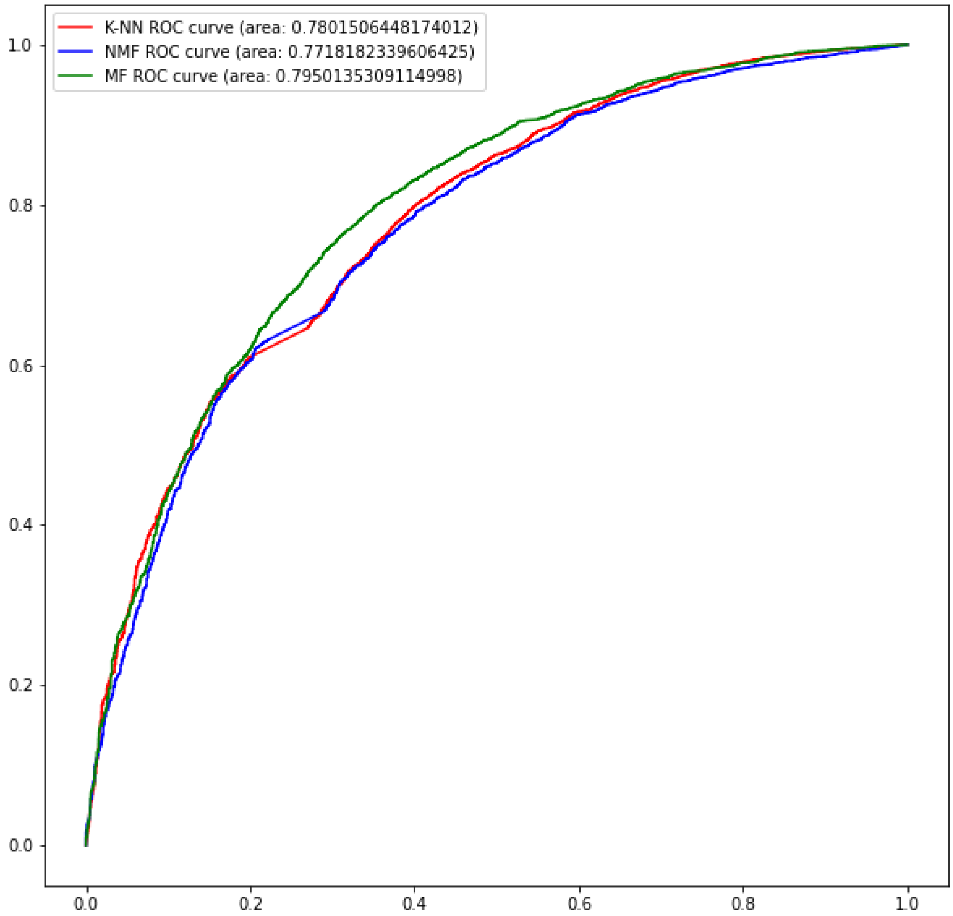


Figure 20: ROC comparison

We set the threshold to be 3 and plotted the ROC curves for the k-NN,NMF, and MF with bias based collaborative filters in the same figure, as shown in the Fig. 20.

As shown in the figure. 20, the AUCs for K-NN, NMF and MF are:

$$AUC(K-nn) = 0.78015$$

$$AUC(NMF) = 0.771818$$

$$AUC(MF) = 0.795013$$

Question 13

The mathematical definition of precision and recall are given below:

$$Percision(t) = \frac{|S(t)| \in G}{|S(t)|}$$

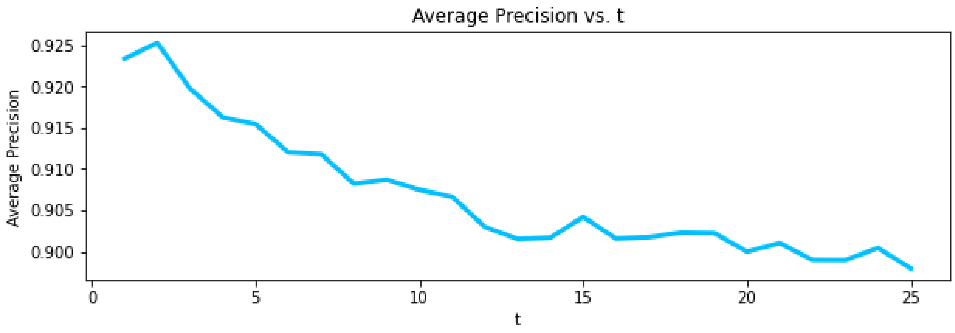
$$Recall(t) = \frac{|S(t)| \in G}{|G|}$$

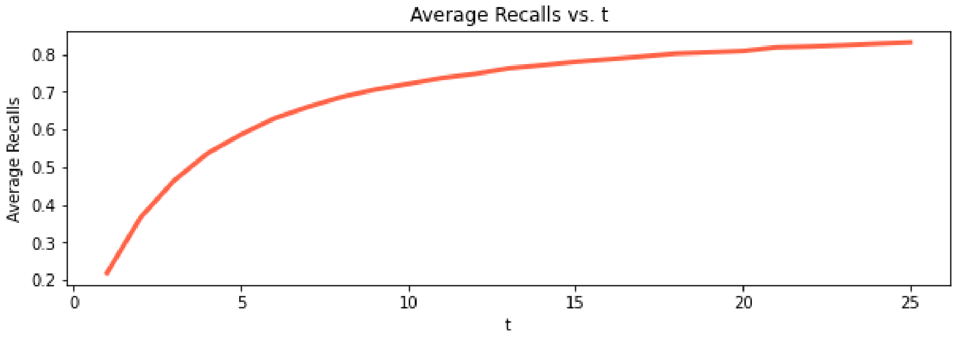
Based on the definition, precision means the percentage of items recommended to the user that the user actually likes, over the total number of recommended items. Whereas recall means the precentage of items recommended to the user that the user actually likes, over the total number of user-like items.

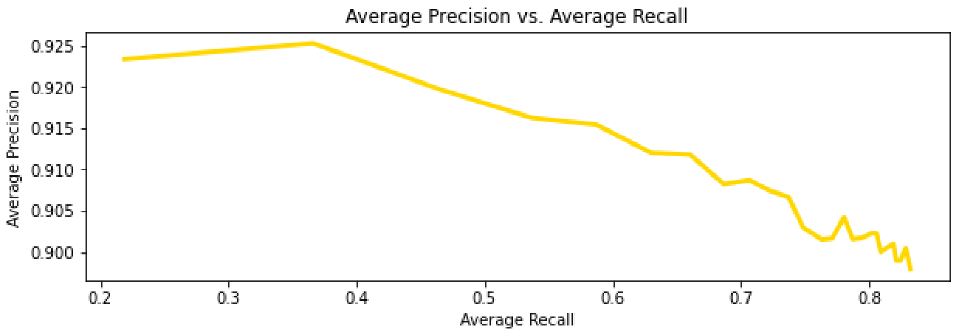
Question 14

Plots

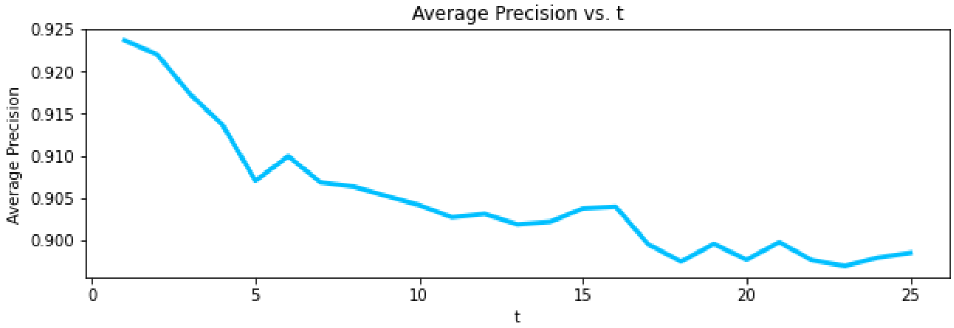
For k-NN, the plots are as follows:

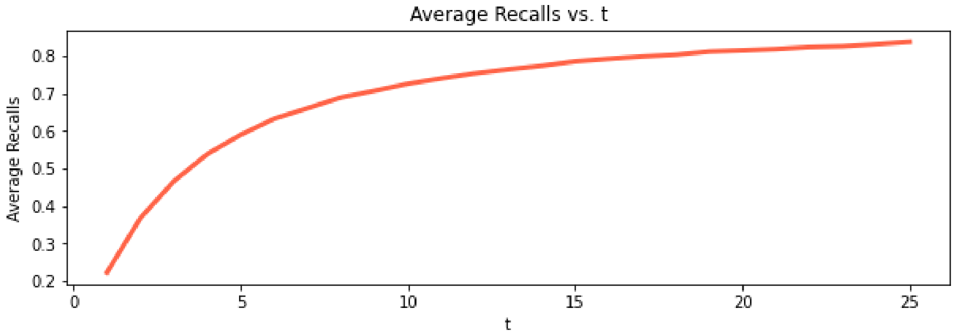


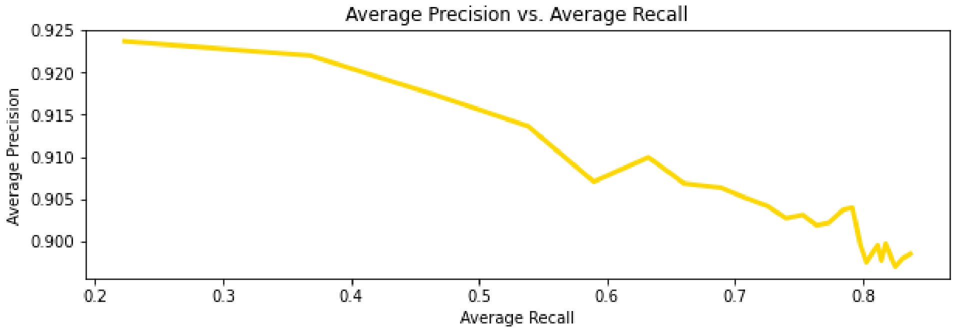




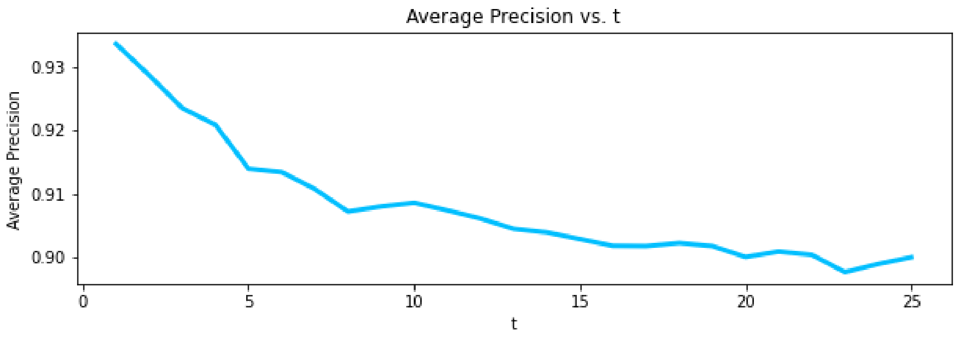
For NMF, the plots are as follows:

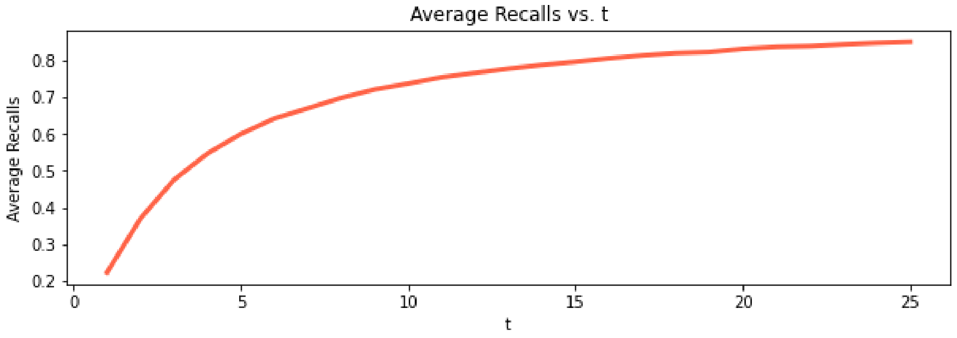


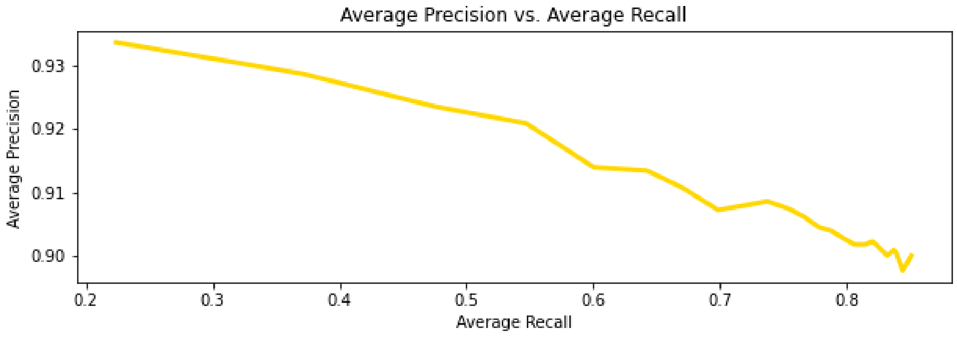




For MF, the plots are as follows:


Average Precision vs. t
Average Precision
0.93
0.92
0.91
0.90
0
5
10
15
20
25
t


Average Recalls vs. t
Average Recalls
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0
5
10
15
20
25
t


Average Precision vs. Average Recall
Average Precision
0.93
0.92
0.91
0.90
0.2
0.3
0.4
0.5
0.6
0.7
0.8
Average Recall

Plot precision-recall curves

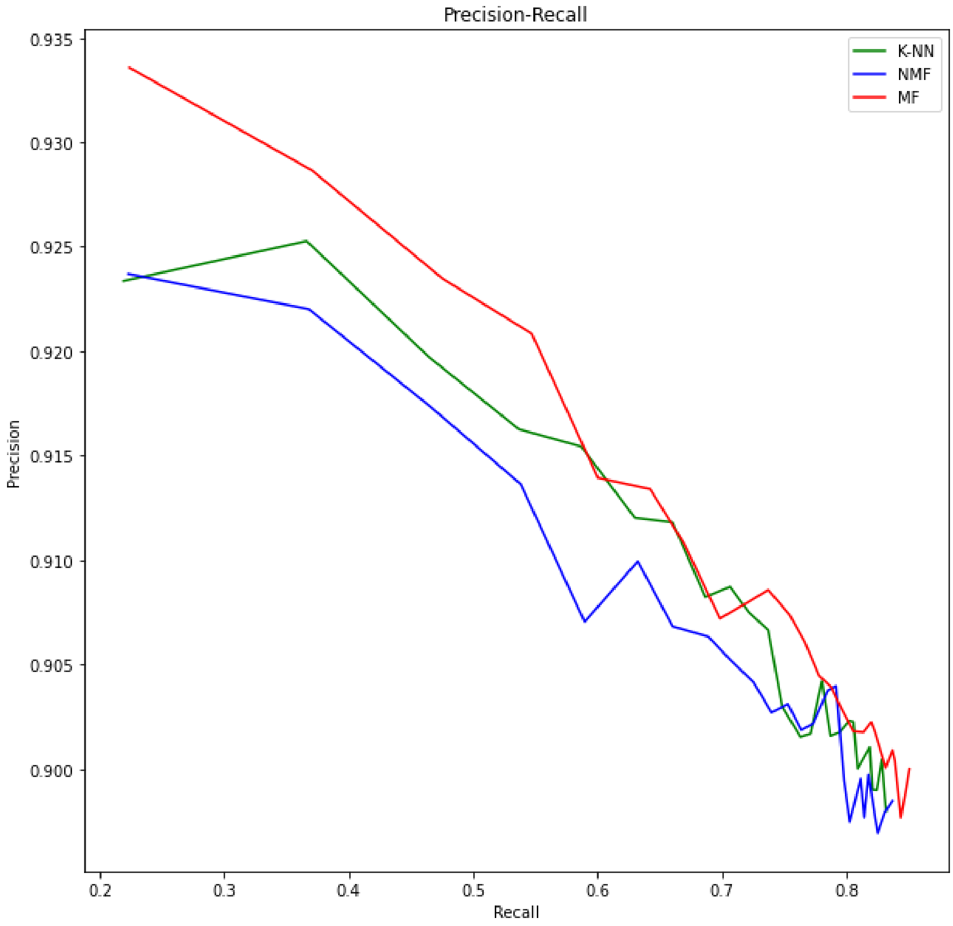


Figure 21: precision-recall curves