CS 6001 - Applied Spatial and Temporal Data Analysis

Homework 4

Weerdhawal Chowgule

RECOMMENDER SYSTEM

Introduction :

Recommender systems are among the most fun and profitable applications of data science in the big data world. Recommender systems aim to predict the rating that a user will give for an item. Training data corresponding to the historical search, browse, purchase, and customer feedback patterns of your customers can be converted into golden opportunities for ROI (*i.e.,* Return on Innovation and Investment). The predictive analytics tools of data science yield a bonanza of mechanisms to engage your customers and enrich their customer experience. What better loyalty program can there be if not the one that offers the customer what they want before they ask and sometimes, even before they think of it for themselves.

Problem Statement:

To constructs a Recommender System for Restaurants.

Performance Metrics:

i. MAE:

In statistics, the **mean absolute error (MAE)** is a quantity used to measure how close forecasts or predictions are to the eventual outcomes. The mean absolute error is given by

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n} |f_i - y_i| = \frac{1}{n}\sum_{i=1}^{n} |e_i| .$$

As the name suggests, the mean absolute error is an average of the absolute errors **|ei|=|fi - yi|**, where **fi** is the prediction and **yi** the true value. Note that alternative formulations may include relative frequencies as weight factors.

The mean absolute error used the same scale as the data being measured. This is known as a scale-dependent accuracy measure and therefore cannot be used to make comparisons between series using different scales.
The mean absolute error is a common measure of forecast error in time series analysis.

ii. RMSE:

The **root-mean-square deviation (RMSD)** or **root-mean-square error (RMSE)** is a frequently used measure of the differences between values (sample and population values) predicted by a model or an estimator and the values observed. The RMSD represents the sample standard deviation of the differences between predicted values and observed values. These individual differences are called residuals when the calculations are performed over the data sample that was used for estimation, and are called *prediction errors* when computed out-of-sample. The RMSD serves to aggregate the magnitudes of the errors in predictions for various times into a single measure of predictive power. RMSD is a good measure of accuracy, but only to compare forecasting errors of different models for a variable and not between variables, as it is scale-dependent.

The RMSD is used to compare differences between two things that may vary, neither of which is accepted as the "standard". For example, when measuring the average difference between two time series $x_{1, t}$ and $x_{2,t}$, the formula is

$$\mathrm{RMSD} = \sqrt{\frac{\sum_{t=1}^{n}(x_{1,t} - x_{2,t})^2}{n}}.$$

Algorithms:

i. SVD:

The **singular value decomposition** (**SVD**) is a factorization of a real or complex matrix. It is the generalization of the eigendecomposition of a positive semidefinite normal matrix (for example, a symmetric matrix with positive eigenvalues) to any m x n matrix via an extension of polar decomposition. It has many useful applications in signal processing and statistics.

ii. PMF:

The **Probabilistic Matrix Factorization algorithm (PMF)** model is based on the assumption that users who have rated similar sets of restaurants are likely to have similar preferences. The resulting model is able to generalize considerably better for users with very few ratings.

iii. **NMF:**

Non-negative matrix factorization (**NMF** or **NNMF**), also **non-negative matrix approximation** is a group of algorithms in multivariate analysis and linear algebra where a matrix **V** is factorized into (usually) two matrices **W** and **H**, with the property that all three matrices have no negative elements. This non-negativity makes the resulting matrices easier to inspect. Also, in applications such as processing of audio spectrograms or muscular activity, non-negativity is inherent to the data being considered. Since the problem is not exactly solvable in general, it is commonly approximated numerically.

iv. **UCF:**

User based Collaborative Filtering (UCF)algorithm is a method of making automatic predictions (filtering) about the interests of a user by collecting preferences or taste information from many users (collaborating). The underlying assumption of the collaborative filtering approach is that if a person *A* has the same opinion as a person *B* on an issue, A is more likely to have B's opinion on a different issue than that of a randomly chosen person.

v. **ICF:**

Item based Collaborative Filtering (ICF)algorithm is a method of making automatic predictions (filtering) about the interests of a user by collecting preferences or taste information from many users (collaborating). The underlying assumption of the collaborative filtering approach is that if a person *A* has the same opinion as a person *B* on an issue, A is more likely to have B's opinion on a different issue than that of a randomly chosen person.

Calculation Results:

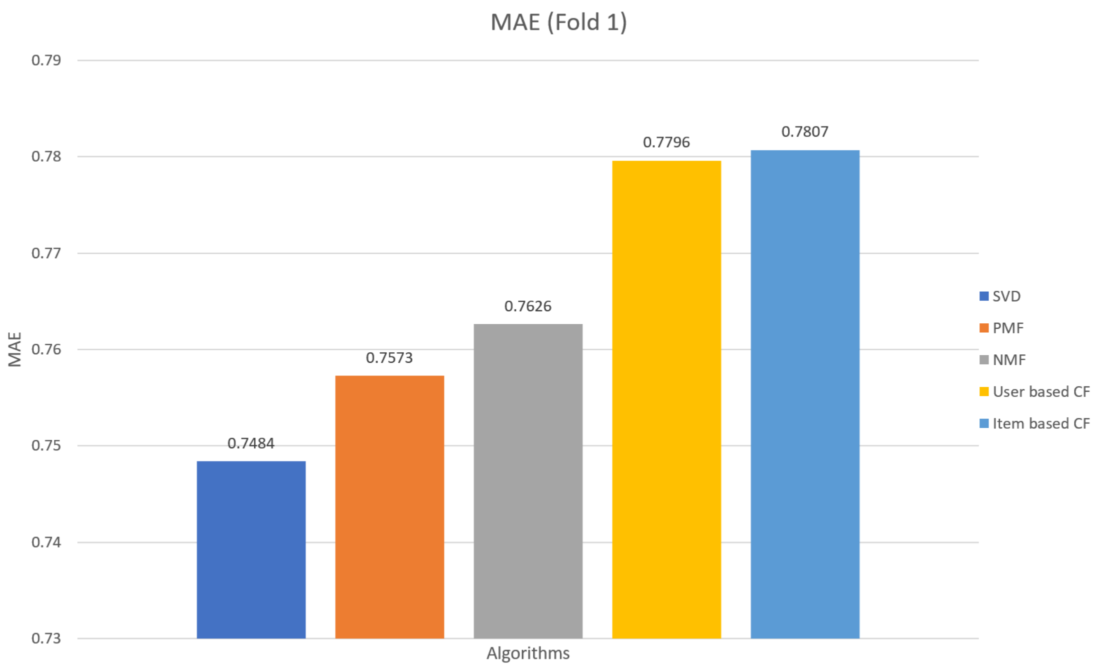
As mentioned in the homework all the different algorithms were applied to dataset provided. 3-fold cross validation also had to be done as a part of this assignment, hence the results of all the folds was calculated and average also was calculated. RMSE and MAE values were also calculated for each fold. The table below shows the corresponding results of all the algorithms in their respective folds. For step 13, cosine, Pearson and Mean-squared distance had to be used as similarity metric which are also calculated and are in the table. In table, Color Green indicates the corresponding lowest error values and Color Red indicates the high error values.

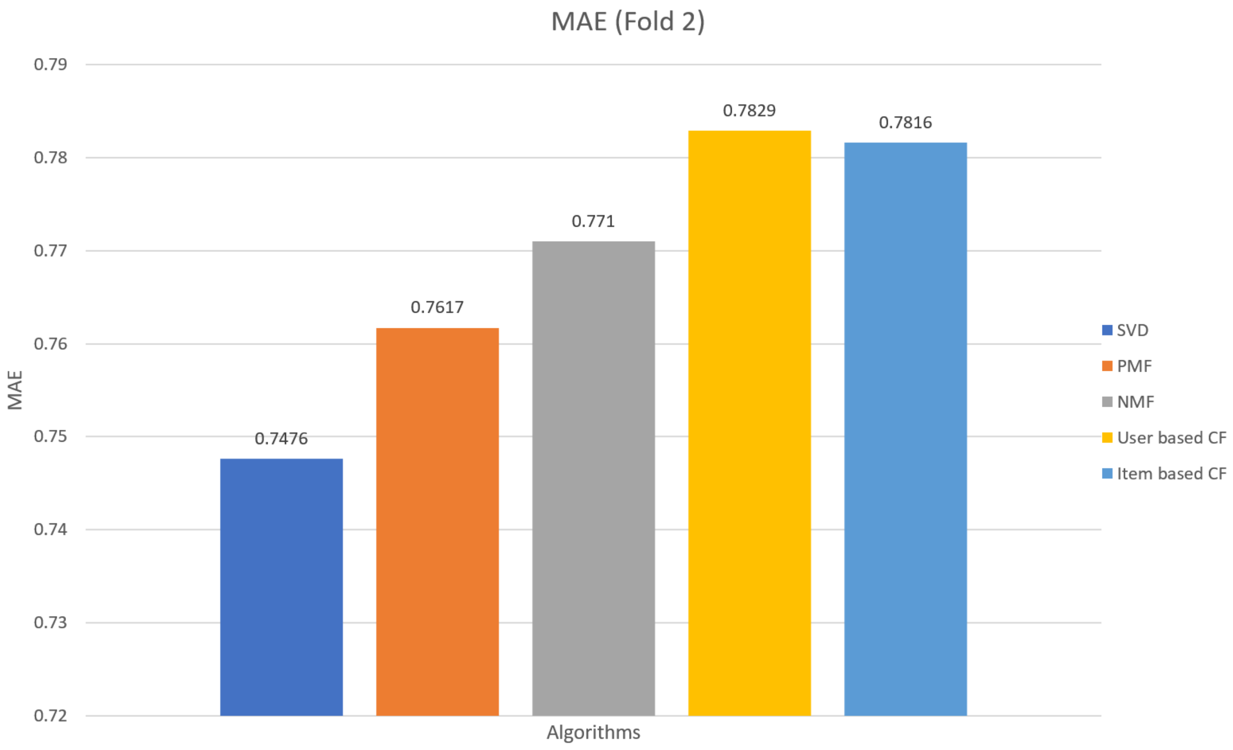
Algorithm/Folds	Fold 1		Fold 2		Fold 3		Mean	
	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
SVD	**0.9466**	**0.7484**	**0.9493**	**0.7476**	**0.9433**	**0.745**	**0.9464**	**0.747**
PMF	0.9614	0.7573	0.9671	0.7617	0.9649	0.7613	0.9645	0.7601
NMF	0.9689	0.7626	0.981	0.771	0.9738	0.7649	0.9745	0.7662
UCF	0.9861	0.7796	0.9917	0.7829	0.9876	0.7808	0.9885	0.7811
ICF	0.9851	0.7807	0.9876	0.7816	0.9852	0.7823	0.986	0.7815
UCF with MSD	0.9861	0.7796	0.9917	0.7829	0.9876	0.7808	0.9885	0.7811
UCF with cosine	1.0188	0.8065	1.0236	0.8098	1.0206	0.8084	1.021	0.8082
UCF with pearson	1.0192	0.809	1.0204	0.8095	1.0213	0.8105	1.0203	0.8096
ICF with MSD	0.9851	0.7807	0.9876	0.7816	0.9852	0.7823	0.986	0.7815
ICF with cosine	1.0359	0.8224	1.0367	0.8231	**1.355**	0.8243	1.036	0.8232
ICF with pearson	**1.0481**	**0.8374**	**1.0502**	**0.8407**	1.0463	**0.839**	**1.0482**	**0.839**

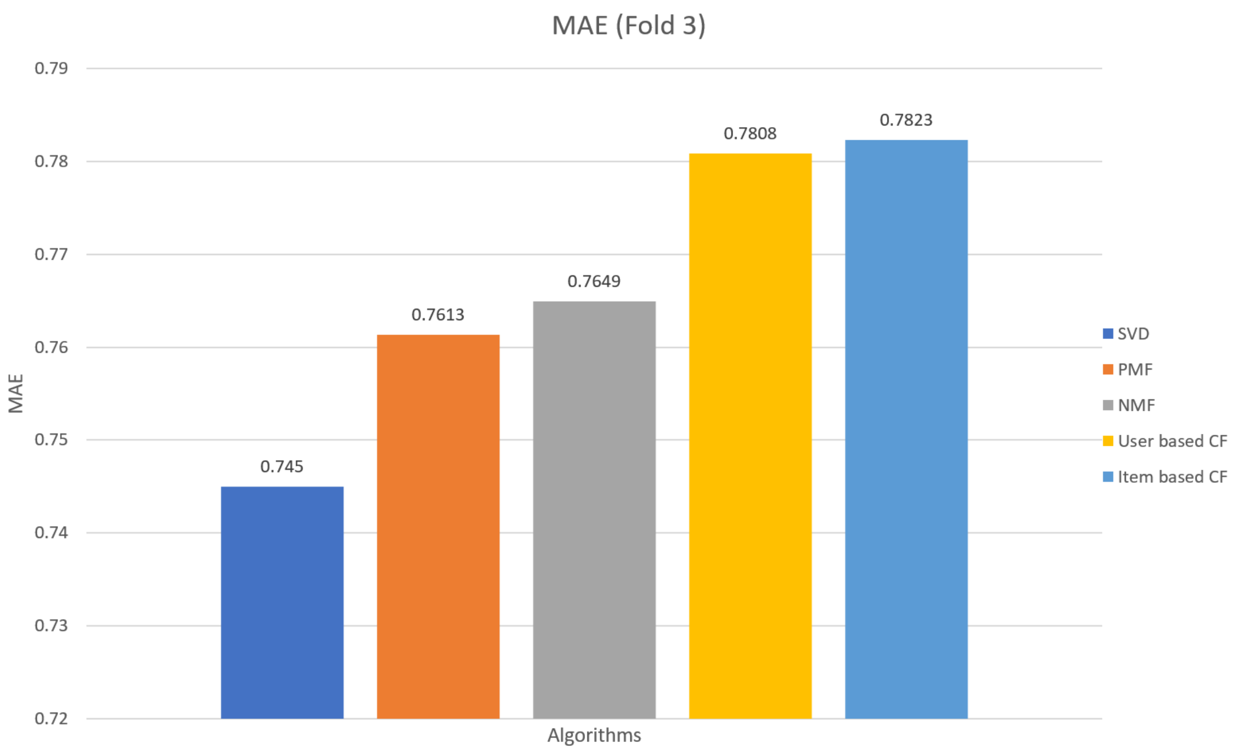
Plots:

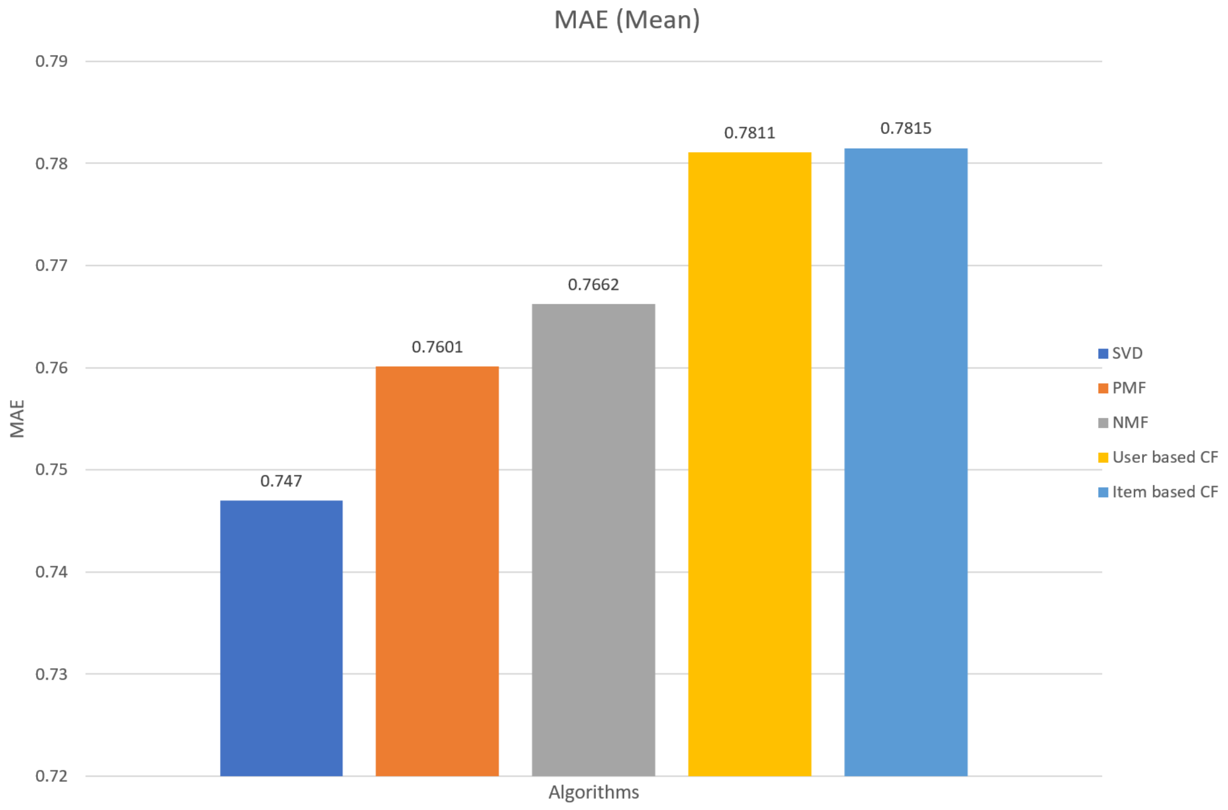
For easy view and better comparison, the values from the table were plotted and this section contains the plots for all of them. The sequence is as follows:

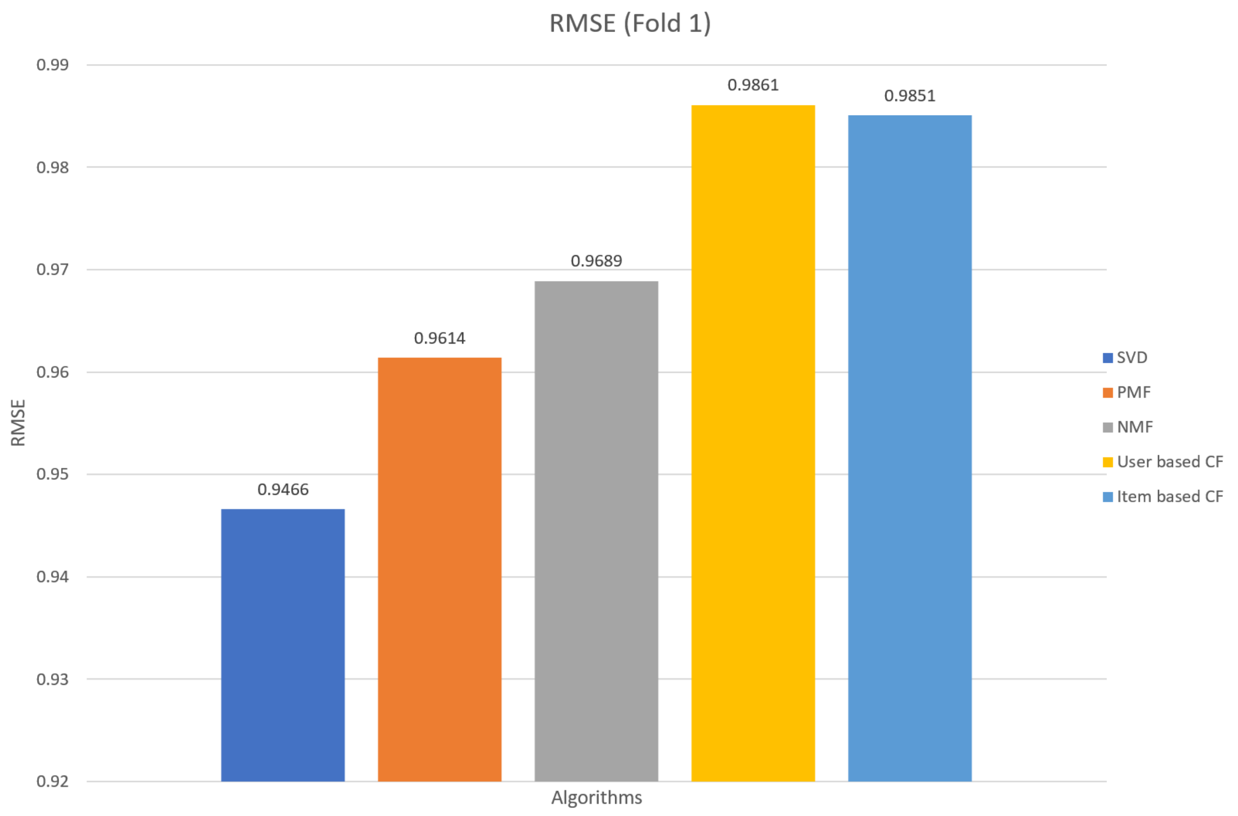
i. MAE (All Folds)
ii. RSME (All Folds)
iii. UCF MAE (All Folds)
iv. UCF RSME (All Folds)
v. ICF MAE (All Folds)
vi. ICF RSME (All Folds)

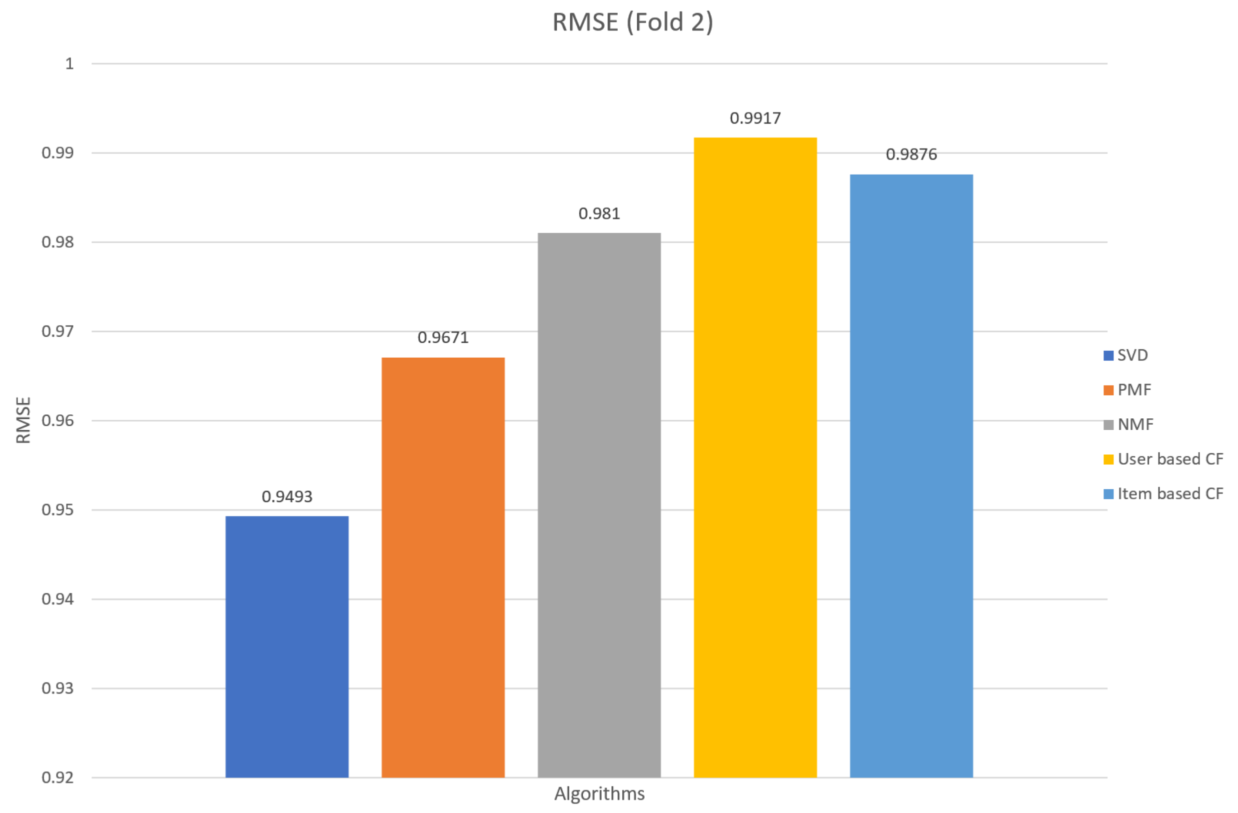

MAE (Fold 1)
0.79
0.78
0.77
0.76
0.75
0.74
0.73
MAE
0.7484
0.7573
0.7626
0.7796
0.7807
Algorithms
SVD
PMF
NMF
User based CF
Item based CF

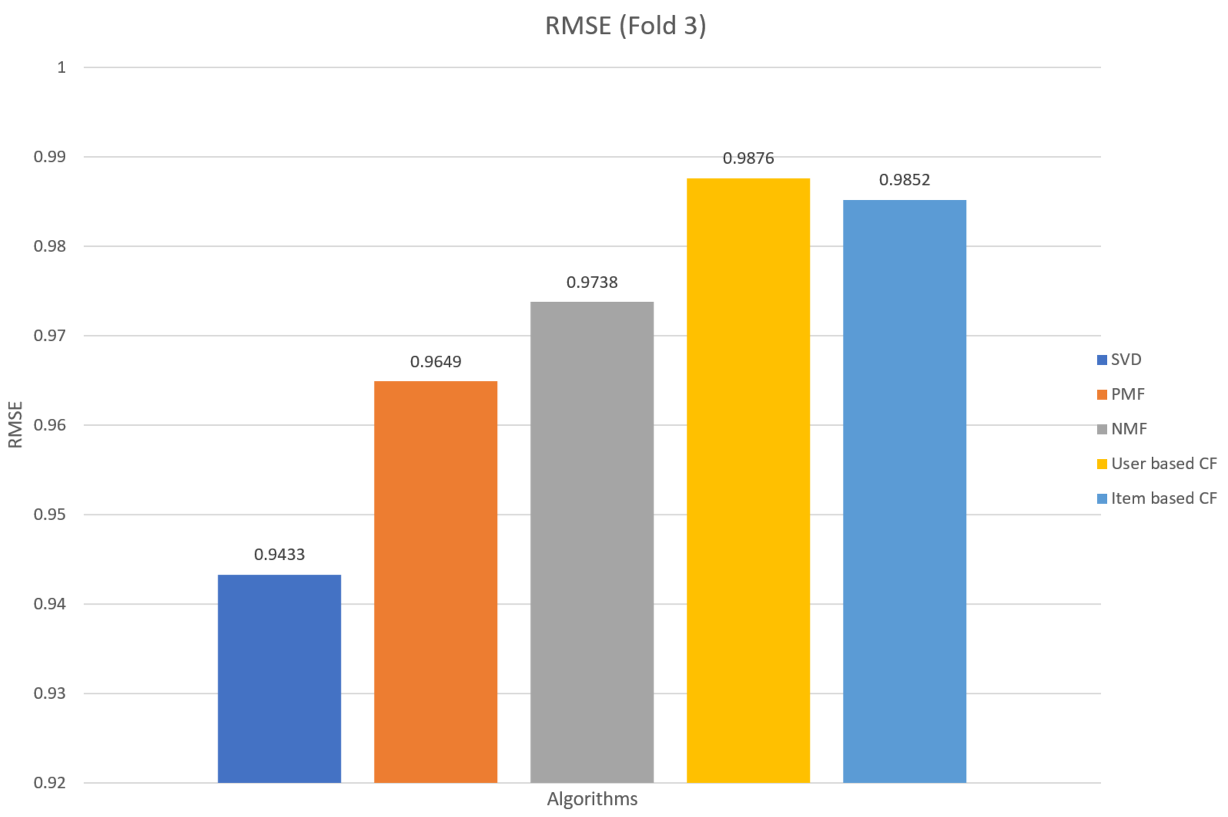

MAE (Fold 2)
0.79
0.78
0.77
0.76
0.75
0.74
0.73
0.72
MAE
0.7476
0.7617
0.771
0.7829
0.7816
Algorithms
SVD
PMF
NMF
User based CF
Item based CF

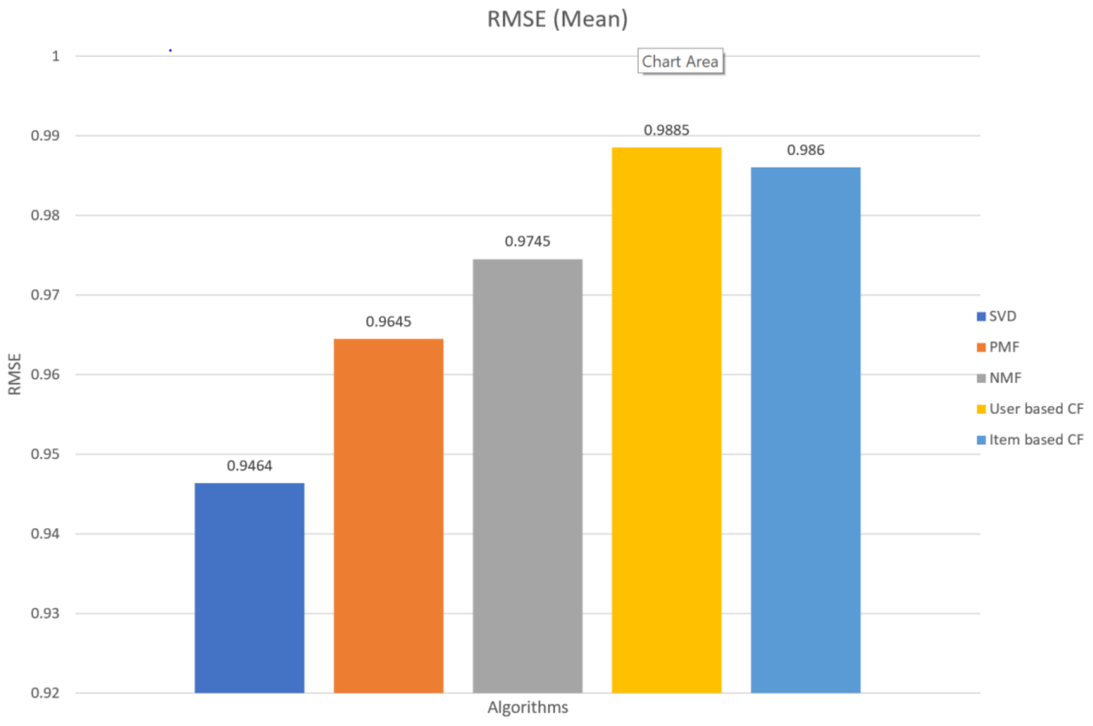

MAE (Fold 3)
0.79
0.78
0.77
0.76
0.75
0.74
0.73
0.72
MAE
0.745
0.7613
0.7649
0.7808
0.7823
Algorithms
SVD
PMF
NMF
User based CF
Item based CF

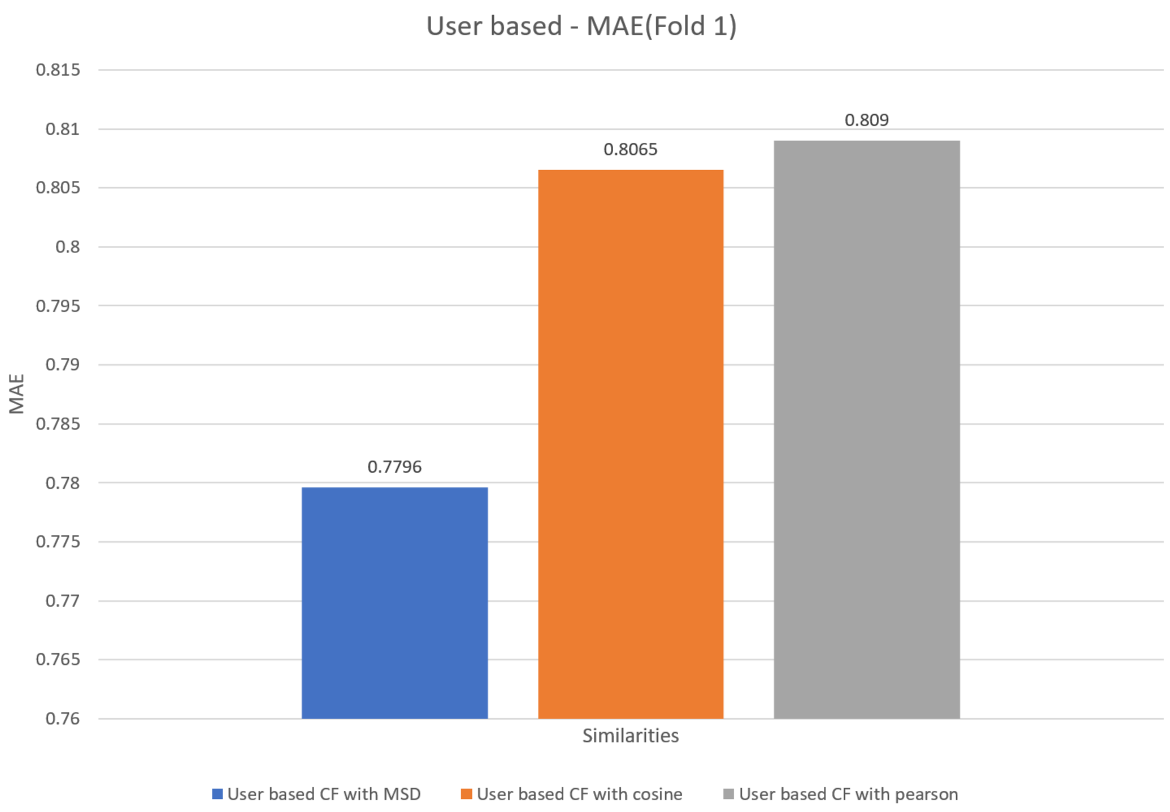

MAE (Mean)
0.79
0.78
0.77
0.76
0.75
0.74
0.73
0.72
MAE
0.747
0.7601
0.7662
0.7811
0.7815
Algorithms
SVD
PMF
NMF
User based CF
Item based CF

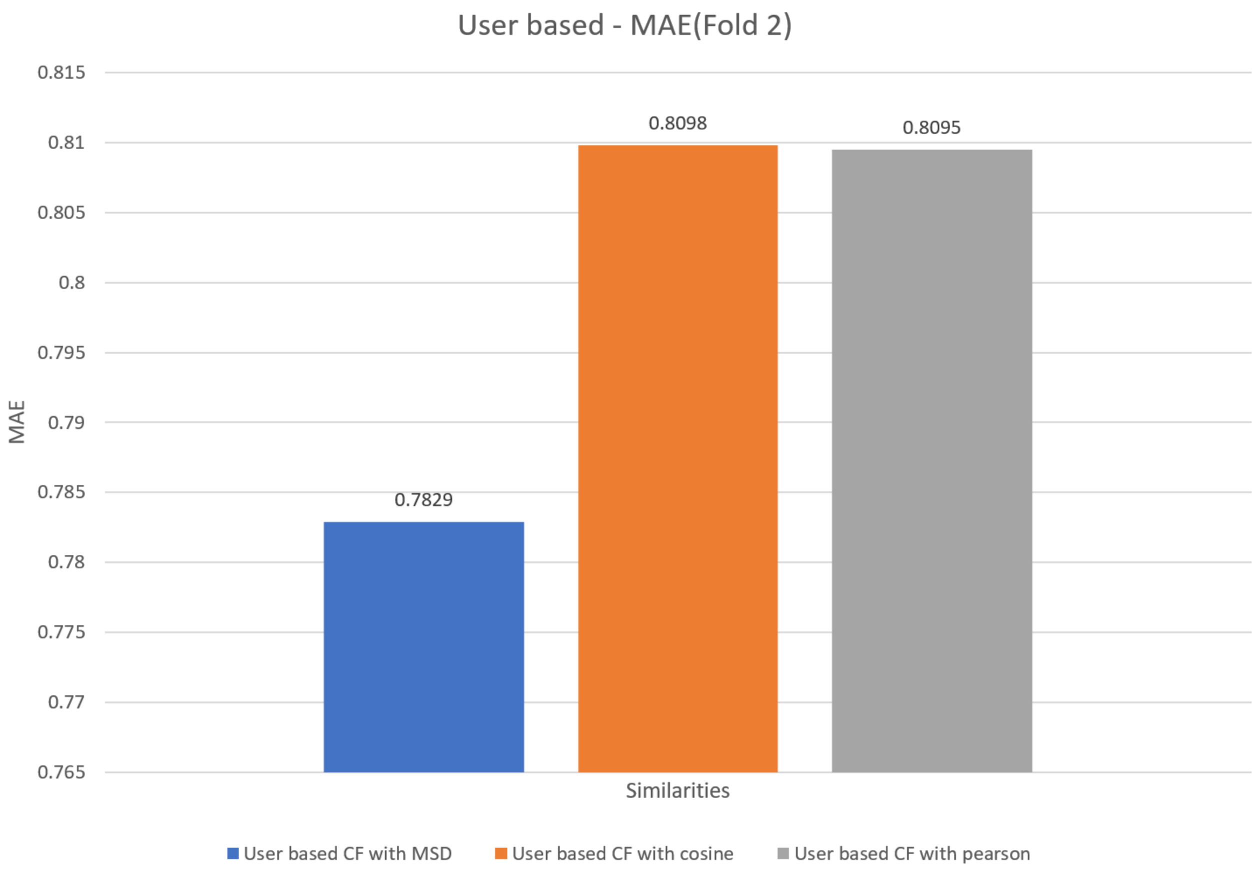

RMSE (Fold 1)
RMSE
0.99
0.98
0.97
0.96
0.95
0.94
0.93
0.92
0.9466
0.9614
0.9689
0.9861
0.9851
Algorithms
■ SVD
■ PMF
■ NMF
■ User based CF
■ Item based CF

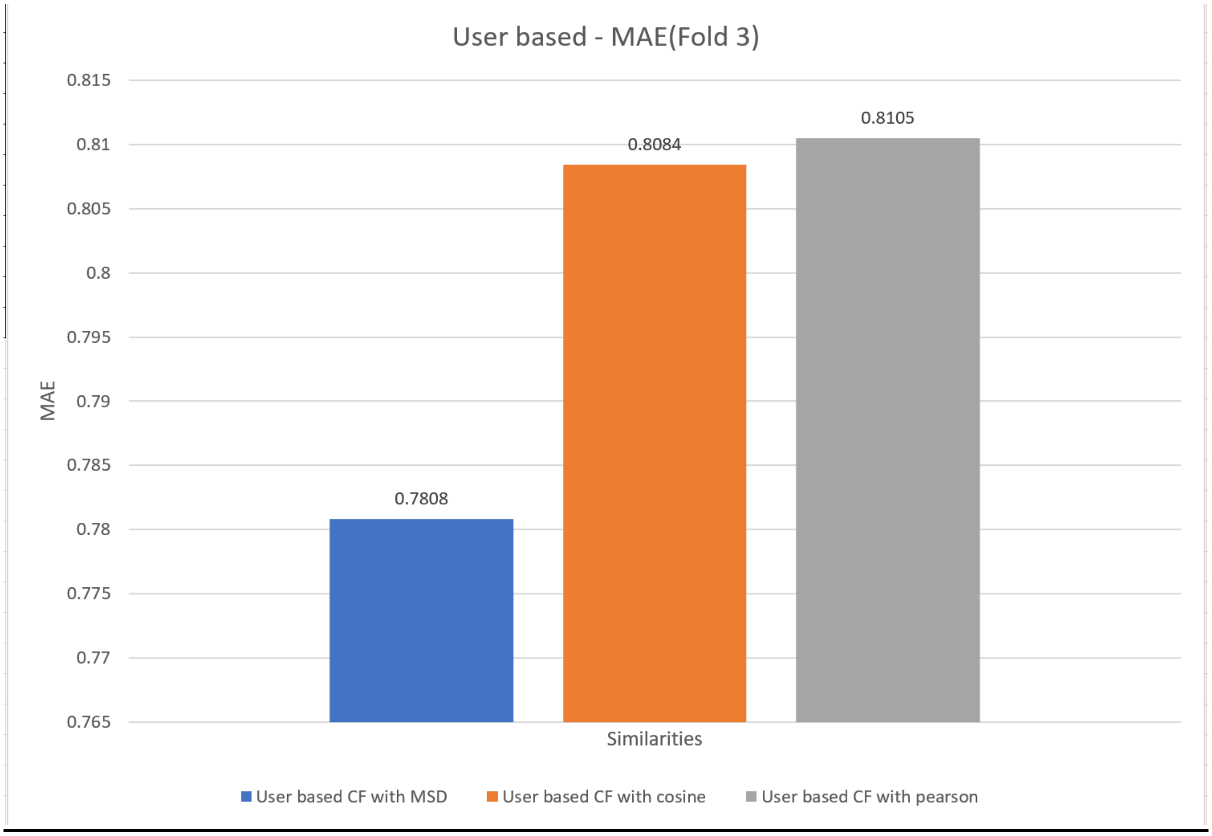

RMSE (Fold 2)
RMSE
1
0.99
0.98
0.97
0.96
0.95
0.94
0.93
0.92
0.9493
0.9671
0.981
0.9917
0.9876
Algorithms
■ SVD
■ PMF
■ NMF
■ User based CF
■ Item based CF

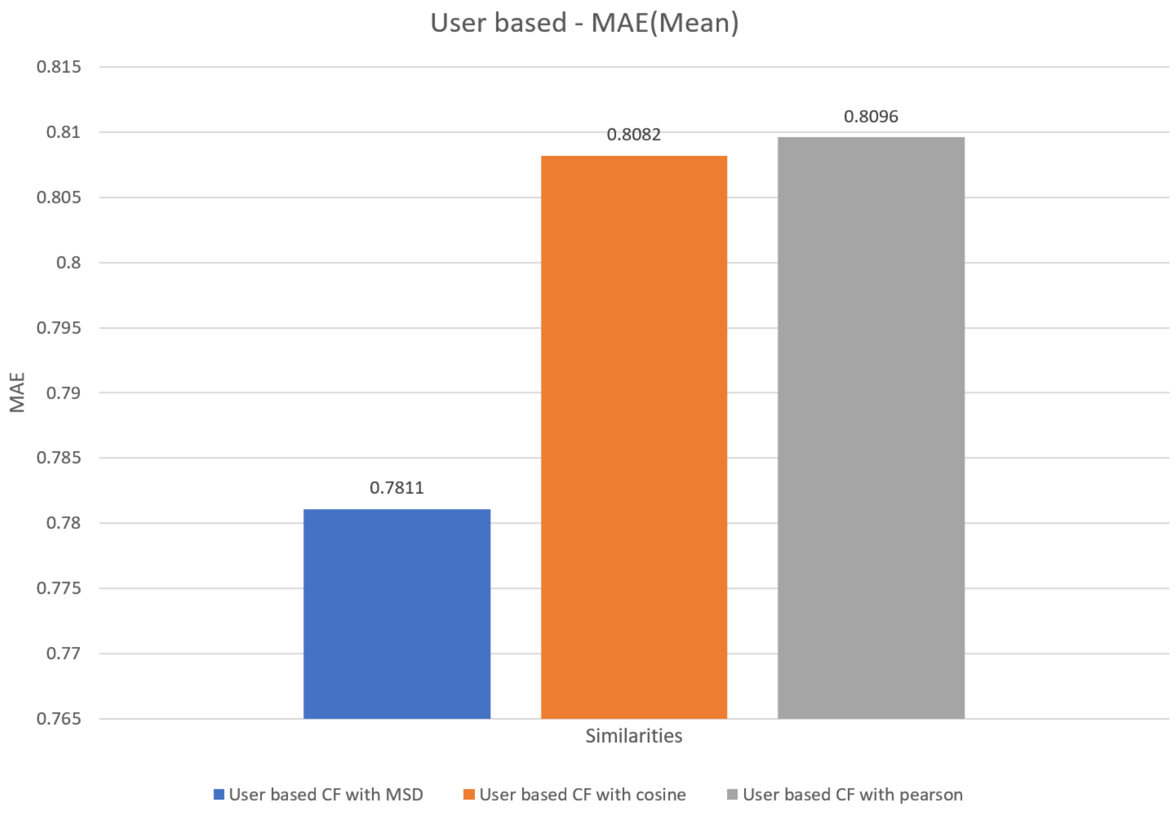

RMSE (Fold 3)
1
0.99
0.98
0.97
0.96
0.95
0.94
0.93
0.92
RMSE
0.9433
0.9649
0.9738
0.9876
0.9852
Algorithms
SVD
PMF
NMF
User based CF
Item based CF

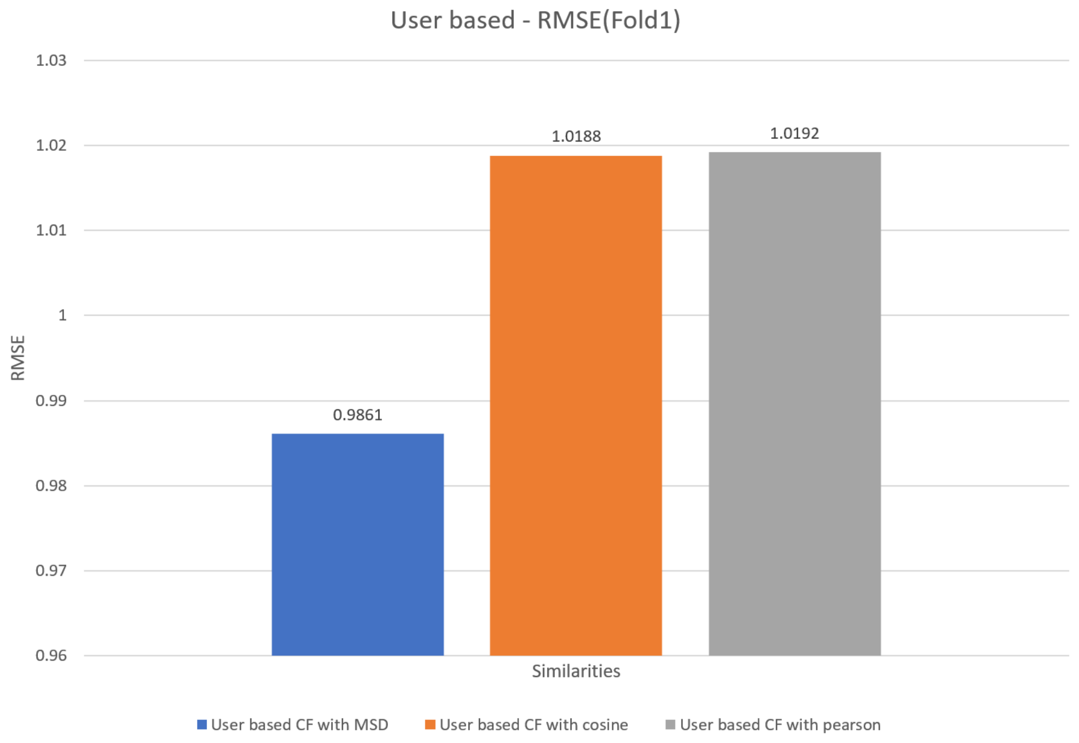

RMSE (Mean)
Chart Area
1
0.99
0.98
0.97
0.96
0.95
0.94
0.93
0.92
RMSE
0.9464
0.9645
0.9745
0.9885
0.986
Algorithms
SVD
PMF
NMF
User based CF
Item based CF

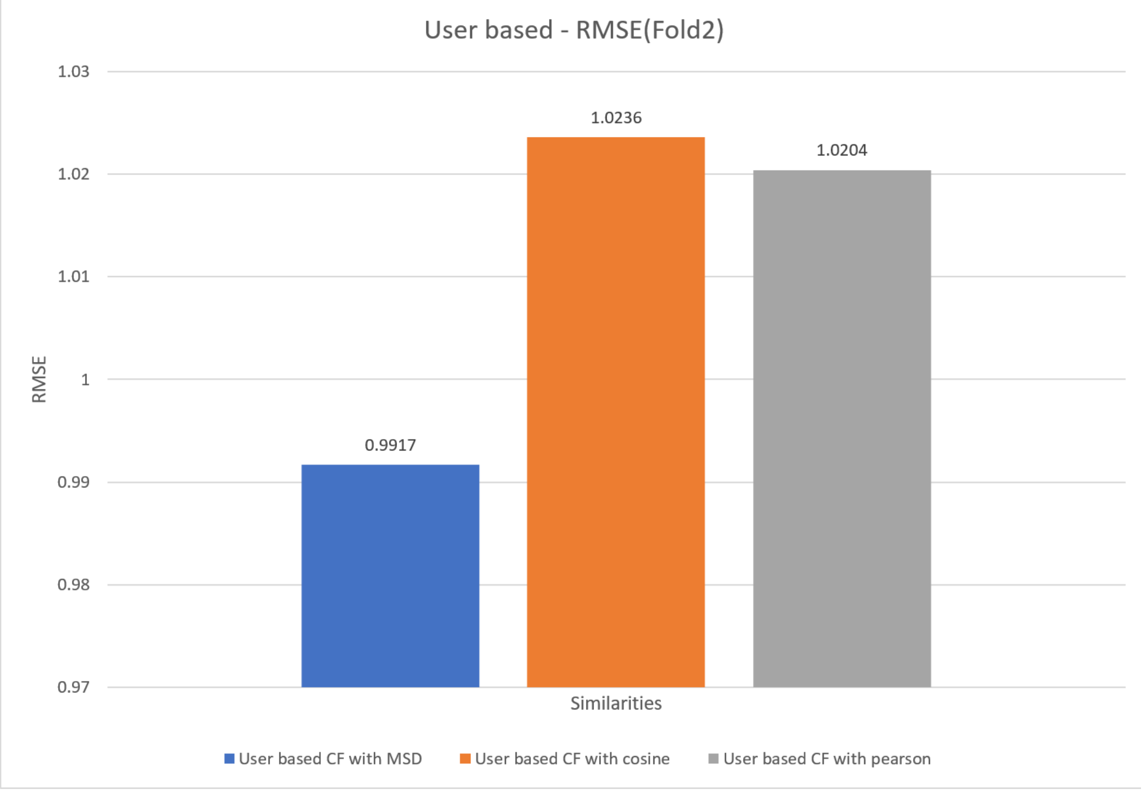

User based - MAE(Fold 1)
MAE
0.815
0.81
0.805
0.8
0.795
0.79
0.785
0.78
0.775
0.77
0.765
0.76
0.7796
0.8065
0.809
Similarities
■ User based CF with MSD
■ User based CF with cosine
■ User based CF with pearson

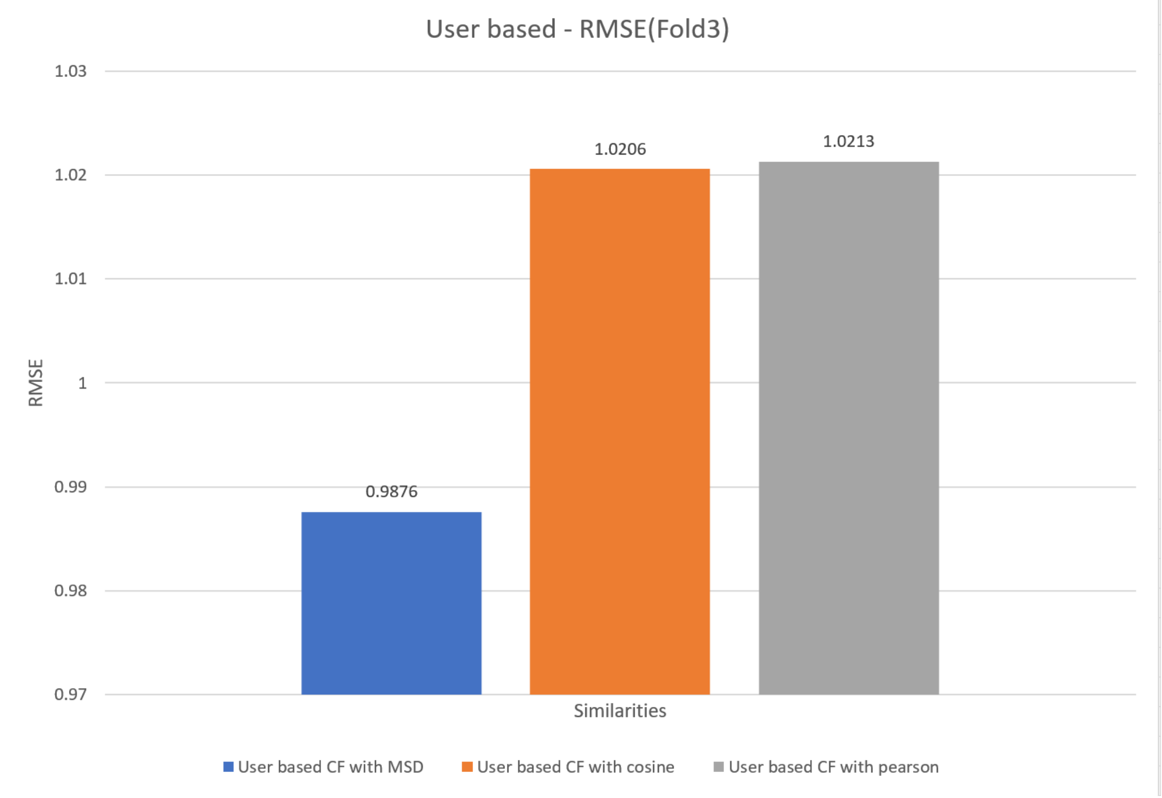

User based - MAE(Fold 2)
MAE
0.815
0.81
0.805
0.8
0.795
0.79
0.785
0.78
0.775
0.77
0.765
0.7829
0.8098
0.8095
Similarities
■ User based CF with MSD
■ User based CF with cosine
■ User based CF with pearson

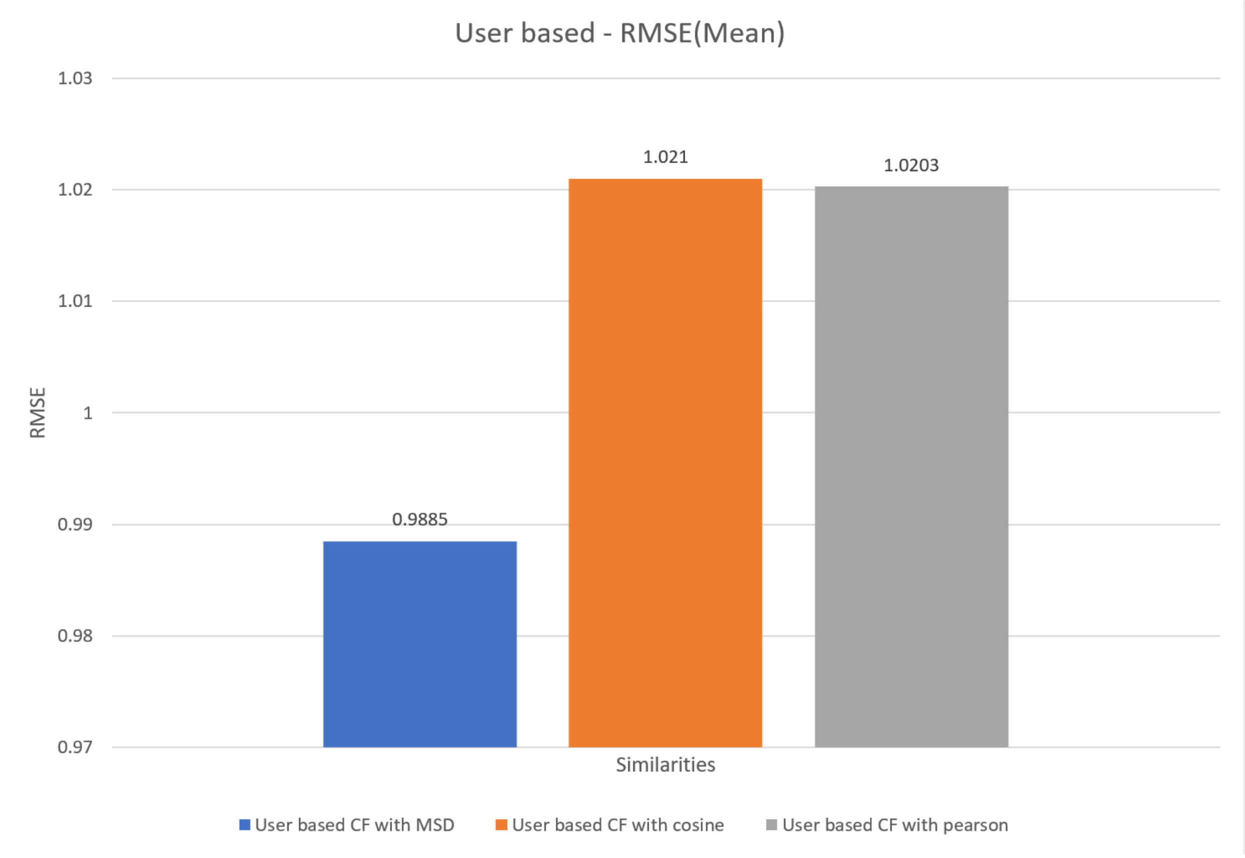

User based - MAE(Fold 3)
MAE
0.815
0.81
0.805
0.8
0.795
0.79
0.785
0.78
0.775
0.77
0.765
0.7808
0.8084
0.8105
Similarities
User based CF with MSD
User based CF with cosine
User based CF with pearson

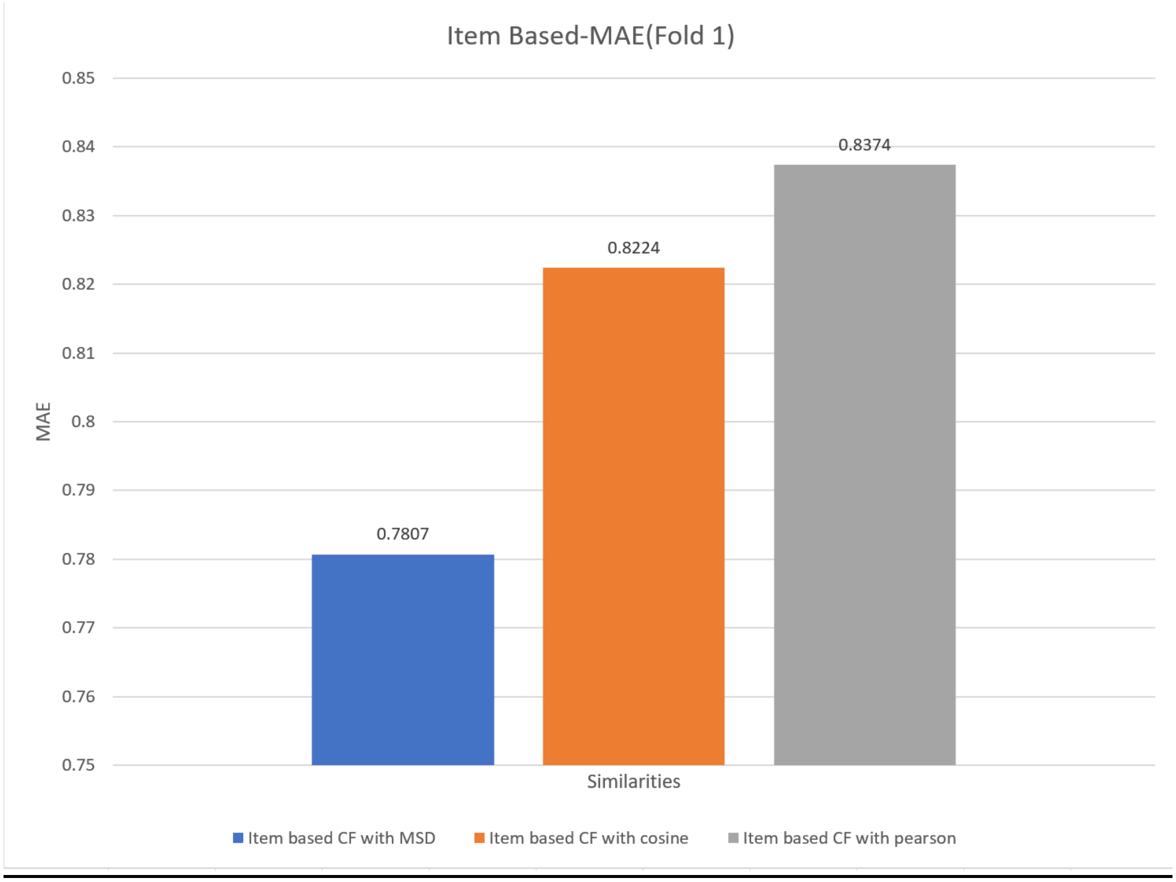

User based - MAE(Mean)
MAE
0.815
0.81
0.805
0.8
0.795
0.79
0.785
0.78
0.775
0.77
0.765
0.7811
0.8082
0.8096
Similarities
■ User based CF with MSD
■ User based CF with cosine
■ User based CF with pearson

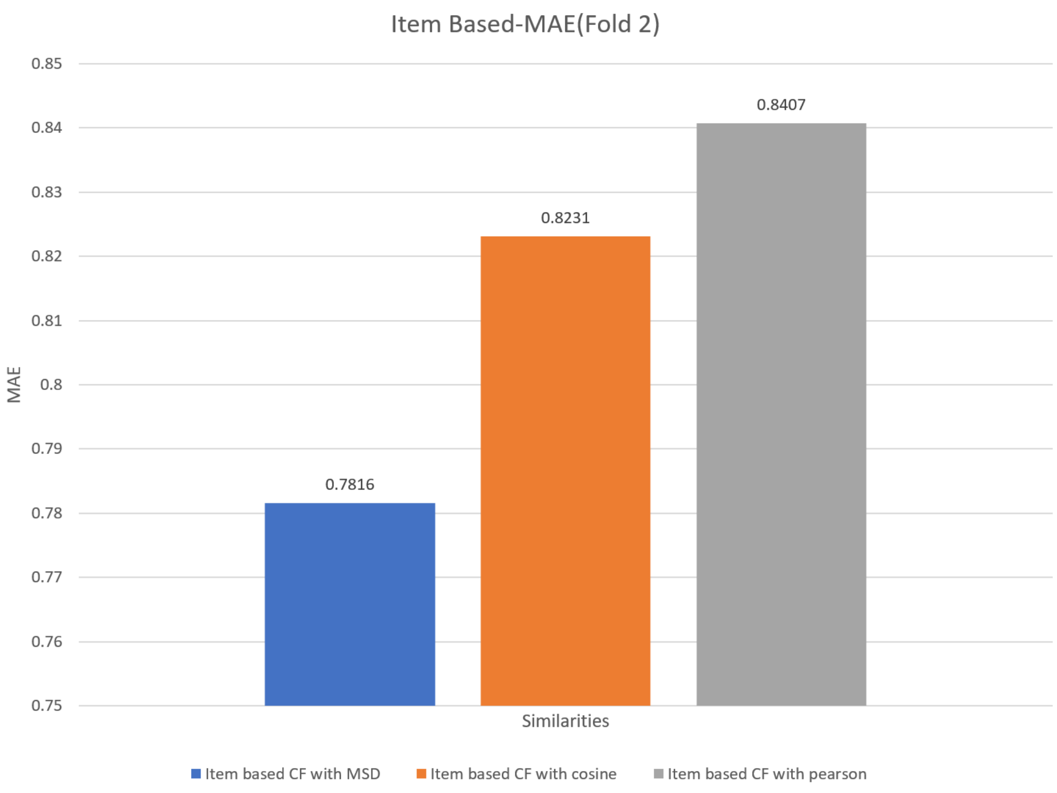

User based - RMSE(Fold1)
RMSE
1.03
1.02
1.01
1
0.99
0.98
0.97
0.96
0.9861
1.0188
1.0192
Similarities
■ User based CF with MSD
■ User based CF with cosine
■ User based CF with pearson

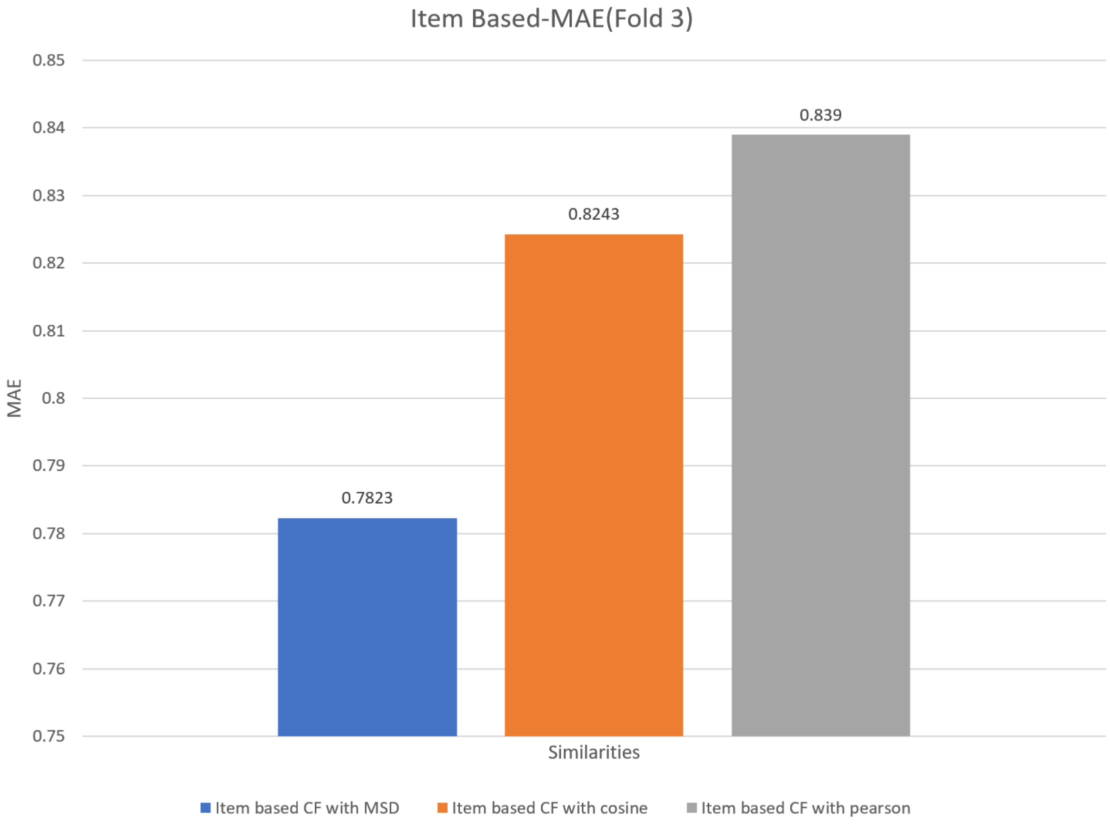

User based - RMSE(Fold2)
RMSE
1.03
1.02
1.01
1
0.99
0.98
0.97
0.9917
1.0236
1.0204
Similarities
■ User based CF with MSD
■ User based CF with cosine
■ User based CF with pearson

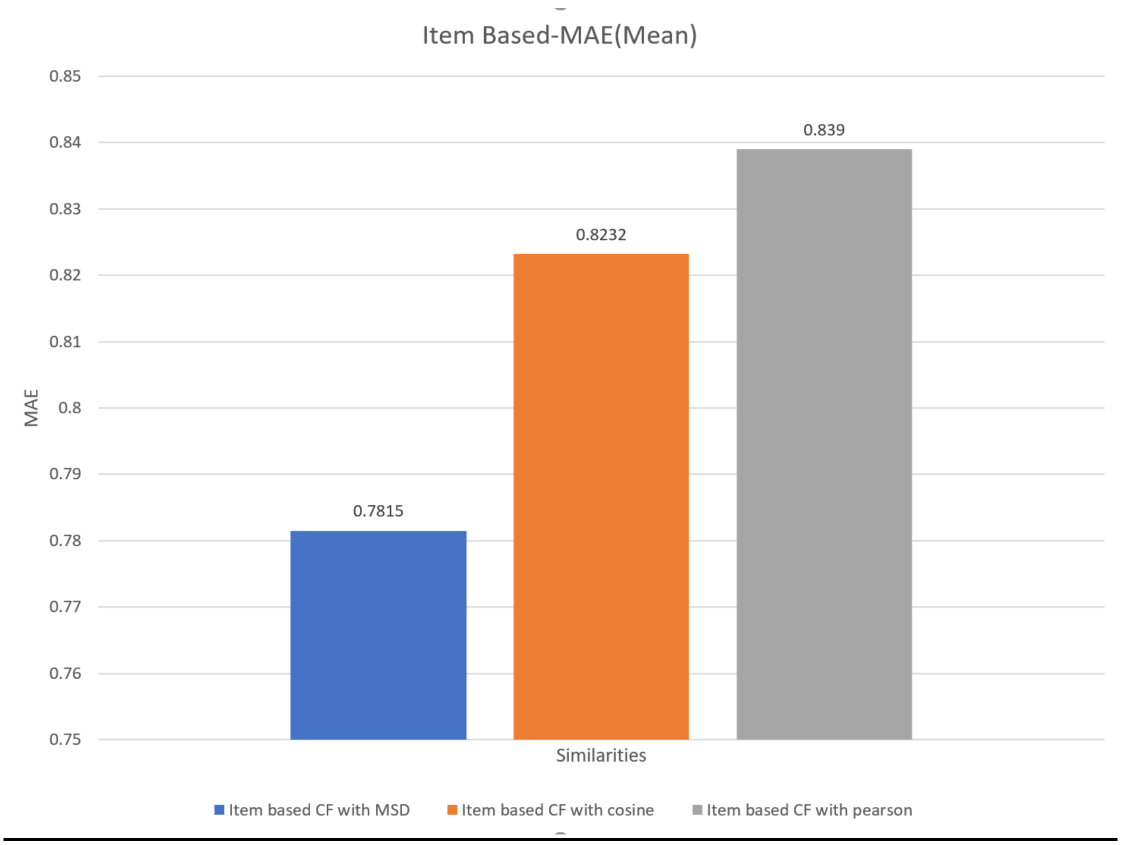

User based - RMSE(Fold3)
RMSE
1.03
1.02
1.01
1
0.99
0.98
0.97
0.9876
1.0206
1.0213
Similarities
■ User based CF with MSD
■ User based CF with cosine
■ User based CF with pearson

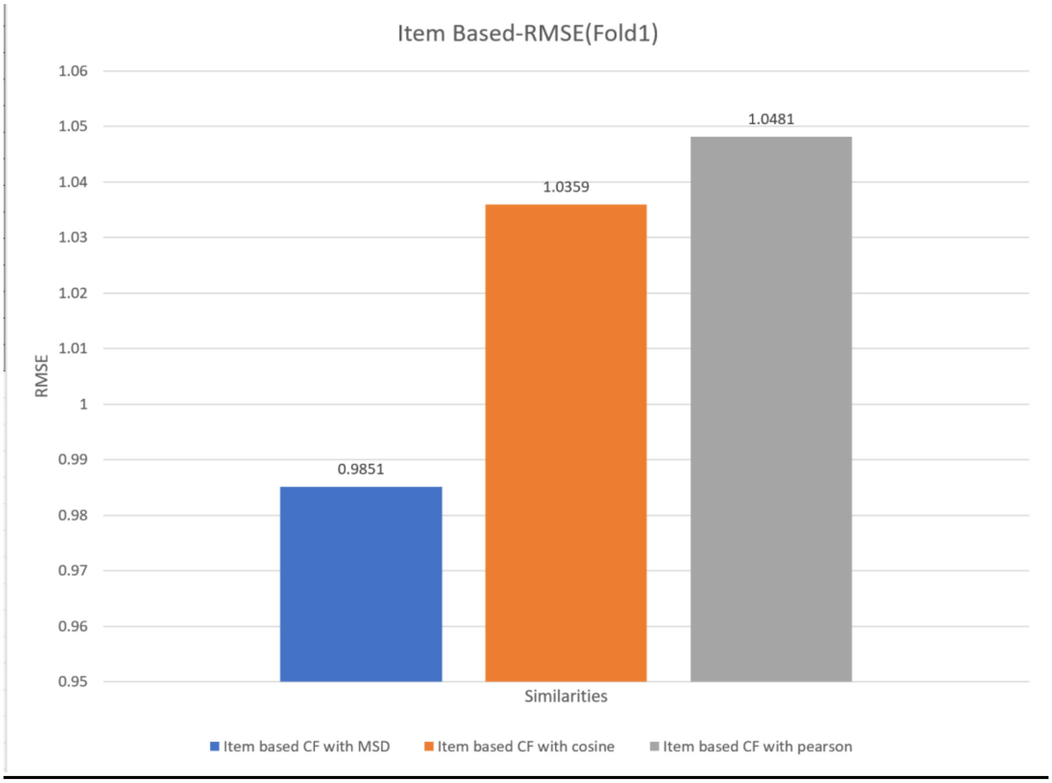

User based - RMSE(Mean)
RMSE
1.03
1.02
1.01
1
0.99
0.98
0.97
0.9885
1.021
1.0203
Similarities
User based CF with MSD
User based CF with cosine
User based CF with pearson

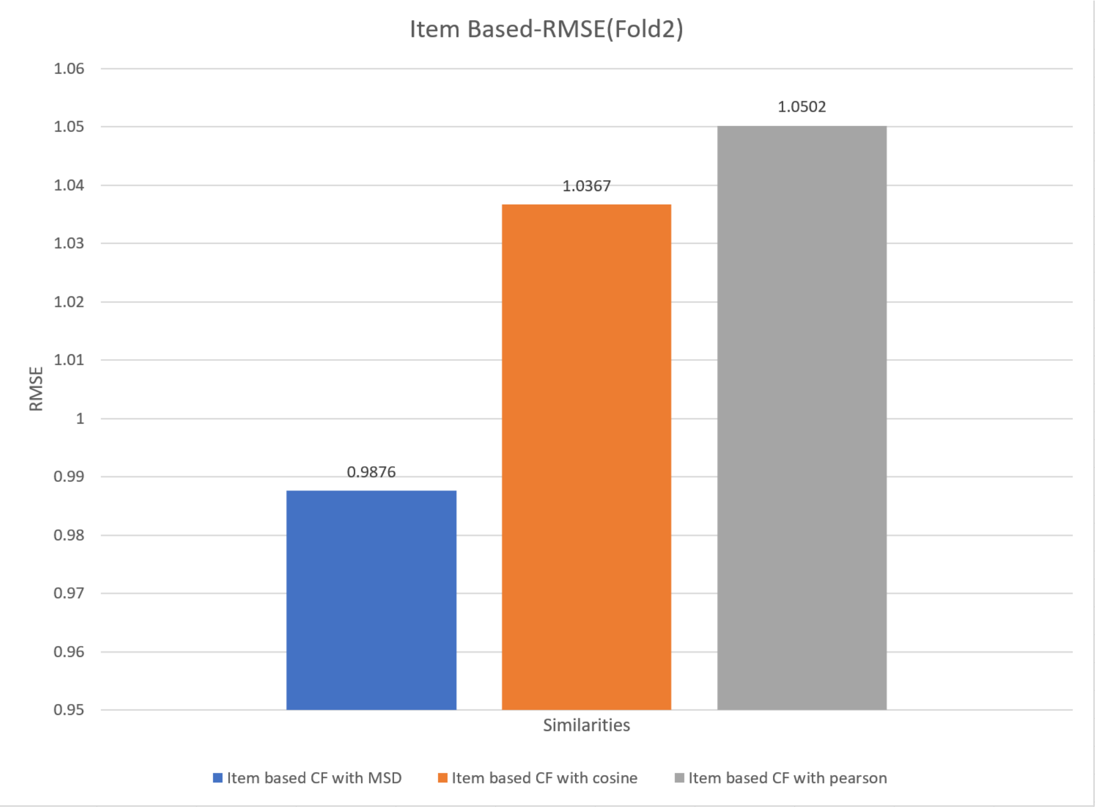

Item Based-MAE(Fold 1)
MAE
0.85
0.84
0.83
0.82
0.81
0.8
0.79
0.78
0.77
0.76
0.75
0.7807
0.8224
0.8374
Similarities
Item based CF with MSD
Item based CF with cosine
Item based CF with pearson

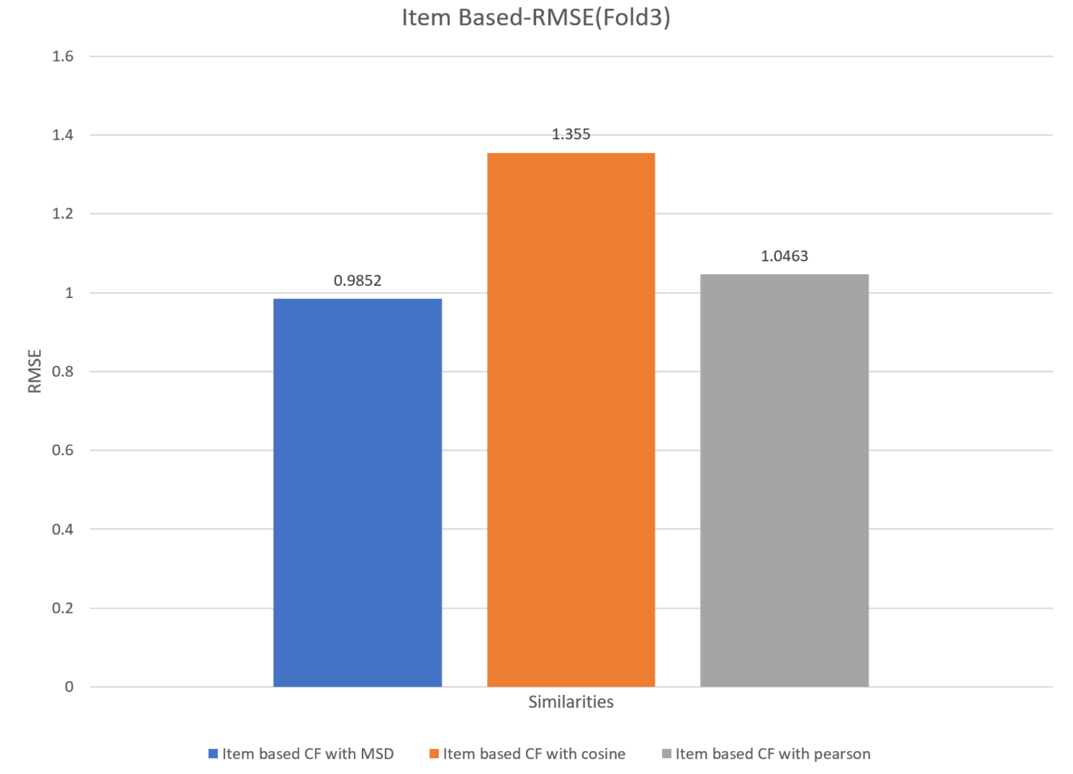

Item Based-MAE(Fold 2)

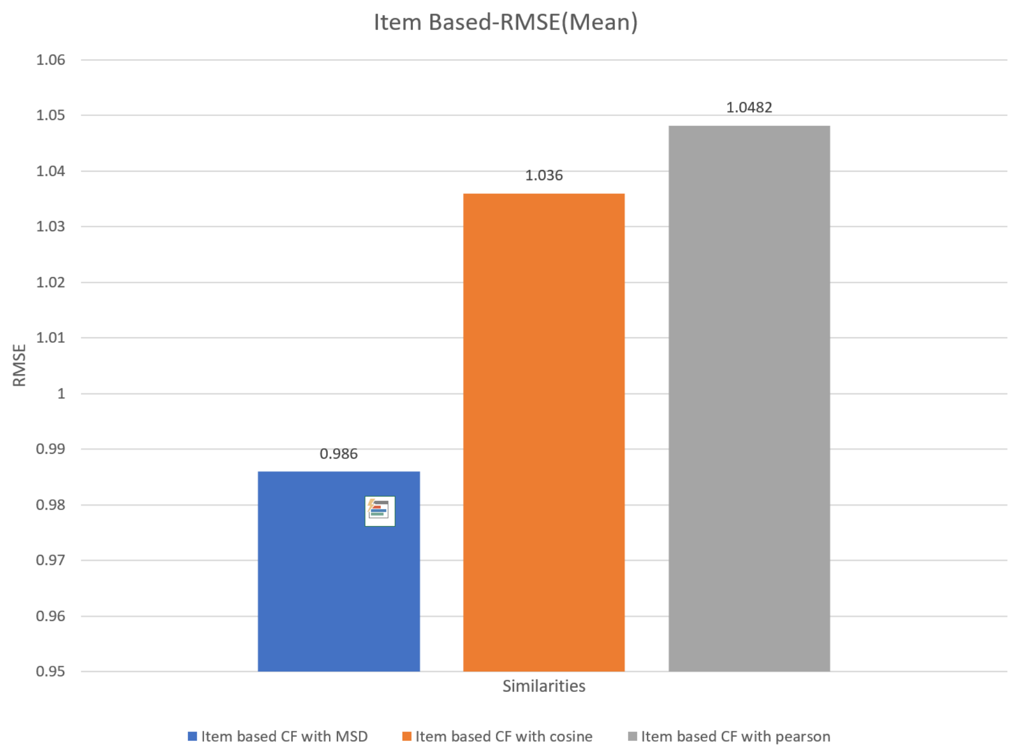

Item Based-MAE(Fold 3)


Item Based-MAE(Mean)
MAE
0.85
0.84
0.83
0.82
0.81
0.8
0.79
0.78
0.77
0.76
0.75
0.7815
0.8232
0.839
Similarities
■ Item based CF with MSD
■ Item based CF with cosine
■ Item based CF with pearson


Item Based-RMSE(Fold1)
RMSE
1.06
1.05
1.04
1.03
1.02
1.01
1
0.99
0.98
0.97
0.96
0.95
0.9851
1.0359
1.0481
Similarities
■ Item based CF with MSD
■ Item based CF with cosine
■ Item based CF with pearson


Item Based-RMSE(Fold2)
1.06
1.05
1.04
1.03
1.02
1.01
1
0.99
0.98
0.97
0.96
0.95
RMSE
0.9876
1.0367
1.0502
Similarities
Item based CF with MSD
Item based CF with cosine
Item based CF with pearson


Item Based-RMSE(Fold3)
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
RMSE
0.9852
1.355
1.0463
Similarities
Item based CF with MSD
Item based CF with cosine
Item based CF with pearson



Comparisons:

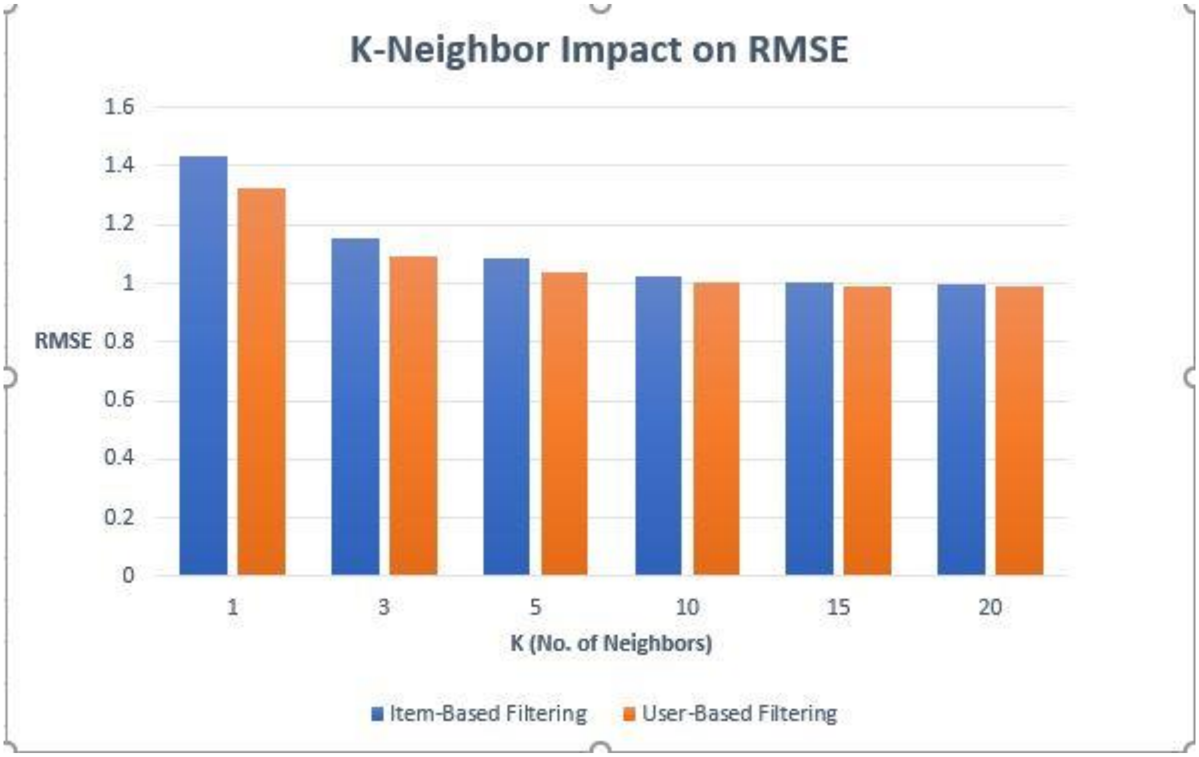
Considering all the algorithms without any distance metrics SVD performed best as it has the lowest scores of the performance metrics RMSE and MSE. UCF and ICF had highest scores of RMSE and MSE which tells that they had the worst performance.

The impact of the three metrics on UCF is kind of consistent with the three metrics in ICF as the trend is an increasing one where, performance-wise MSD>Cosine> Pearson. And obviously MSD performed best in both UCF and ICF.

Step 14: Impact with K Neighbors:

As per the step 14, I applied the KNN algorithm for different values of K (as shown in table and plot) and considered just the RMSE values. We can see that as K increases the RMSE value decrease which in turn means ICF and UCF perform better with higher values of K. At K=20 (considering as the highest value of K) both ICF and UCF performed well.

K / Algorithm	ICF	UCF
1	**1.4342**	**1.3232**
3	1.1537	1.0898
5	1.0843	1.0381
10	1.0232	1.0006
15	1.0038	0.9904
20	**0.9942**	**0.9879**



Final Remark:

SVD algorithm is the best (performance-wise) for the dataset provided.