Movie Recommendation System

Group 2:

Prakruti Joshi (phj15)
Twisha Naik (tn268)
Hari Priya Ponnakanti (hp467)
Dhruv Mundhra (dm1396)

Problem Statement

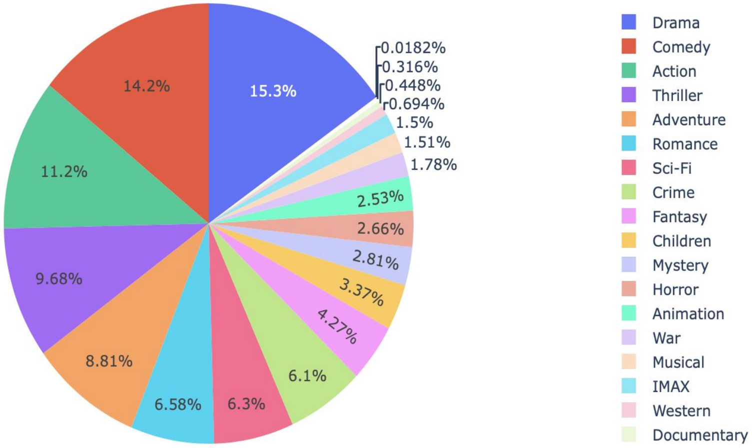
- **Aim:** Build a movie recommendation system based on ‘MovieLens’ dataset.
- We wish to integrate the aspects of personalization of user with the overall features of movie such as genre, popularity etc.

DataSet

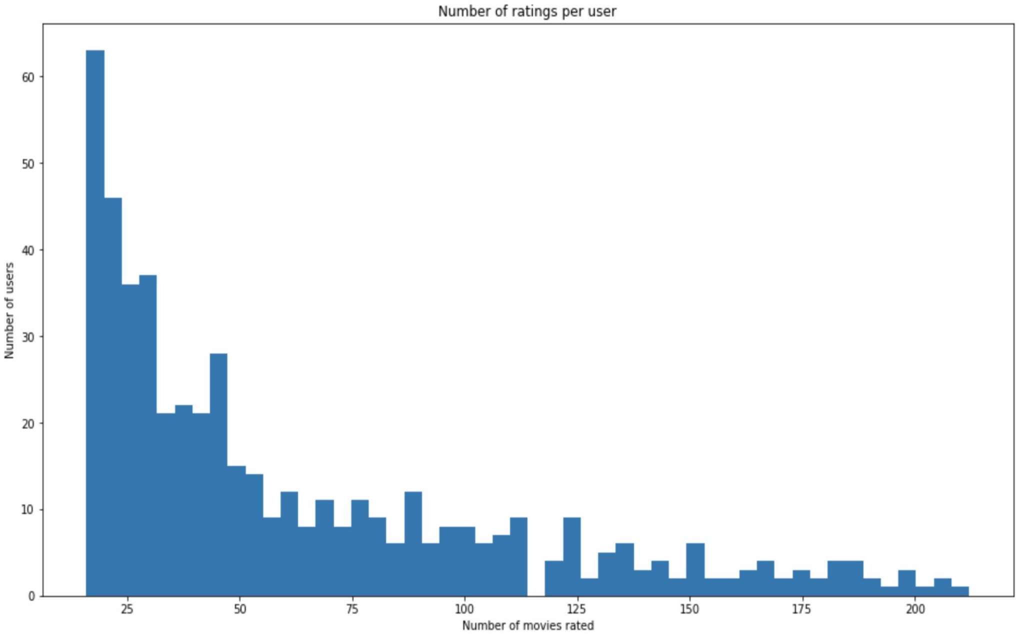
- MovieLens review dataset (ml-latest-small)
  - Ratings: 100k
  - Movies: 9k
  - Users: 600
- Integrated the dataset with IMDB and TMDB data set publically available.
- Split the dataset into 80% training and 20% testing based on the User ID.

Data Analysis

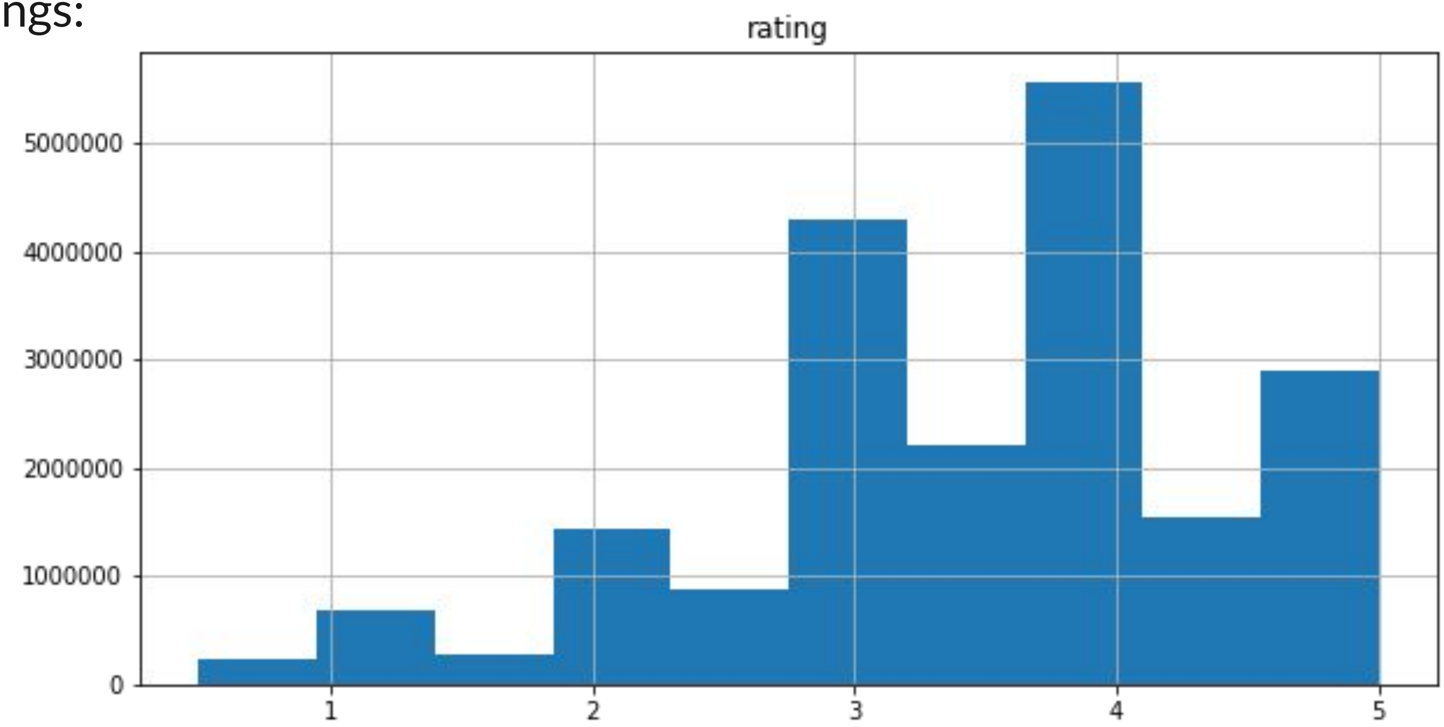
Genre Distribution:



Number of ratings per user:



Histogram of Ratings:

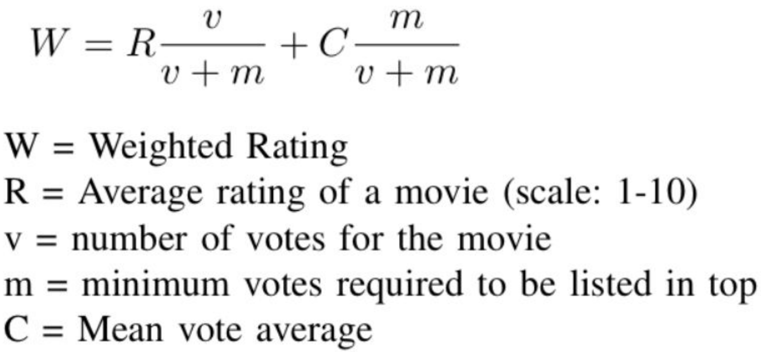


Models

1. Popularity based model
2. Content based model
3. Collaborative Filtering
4. Matrix Factorization method
5. Combined model (SVD + CF)
6. Hybrid model

Popularity Model

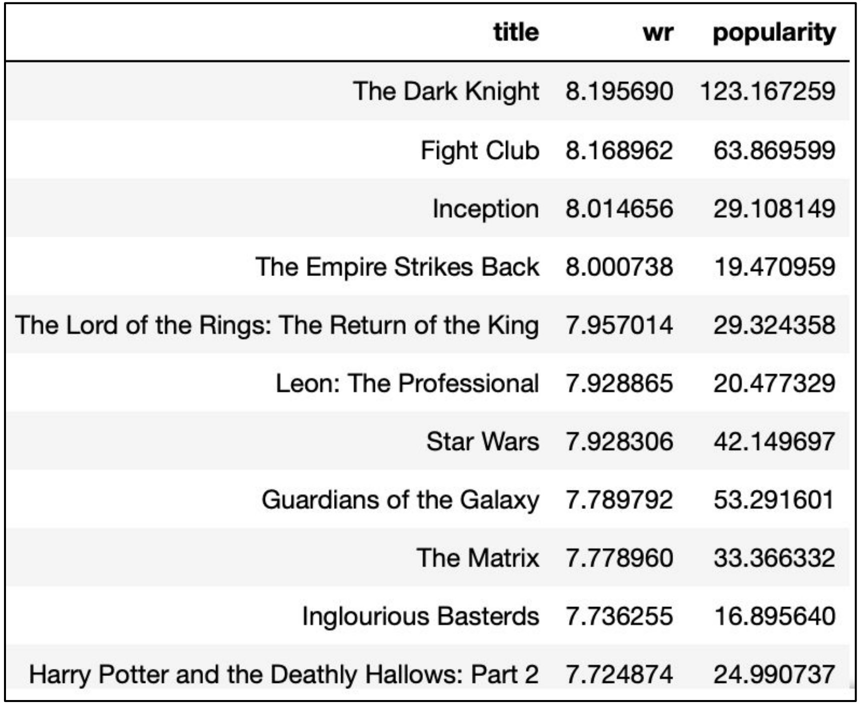
- Genre wise popular movies
- Computed on:
  - Popularity metric from TMDB data
  - Weighted Rating from IMDB



$$W = R\frac{v}{v+m} + C\frac{m}{v+m}$$

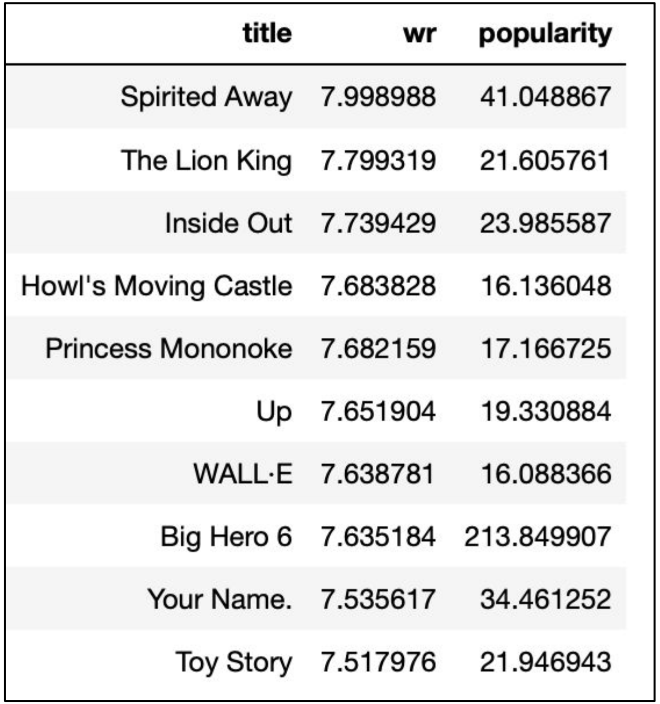
W = Weighted Rating
R = Average rating of a movie (scale: 1-10)
v = number of votes for the movie
m = minimum votes required to be listed in top
C = Mean vote average

Action Movies



title	wr	popularity
The Dark Knight	8.195690	123.167259
Fight Club	8.168962	63.869599
Inception	8.014656	29.108149
The Empire Strikes Back	8.000738	19.470959
The Lord of the Rings: The Return of the King	7.957014	29.324358
Leon: The Professional	7.928865	20.477329
Star Wars	7.928306	42.149697
Guardians of the Galaxy	7.789792	53.291601
The Matrix	7.778960	33.366332
Inglourious Basterds	7.736255	16.895640
Harry Potter and the Deathly Hallows: Part 2	7.724874	24.990737

Animated Movies



title	wr	popularity
Spirited Away	7.998988	41.048867
The Lion King	7.799319	21.605761
Inside Out	7.739429	23.985587
Howl's Moving Castle	7.683828	16.136048
Princess Mononoke	7.682159	17.166725
Up	7.651904	19.330884
WALL·E	7.638781	16.088366
Big Hero 6	7.635184	213.849907
Your Name.	7.535617	34.461252
Toy Story	7.517976	21.946943

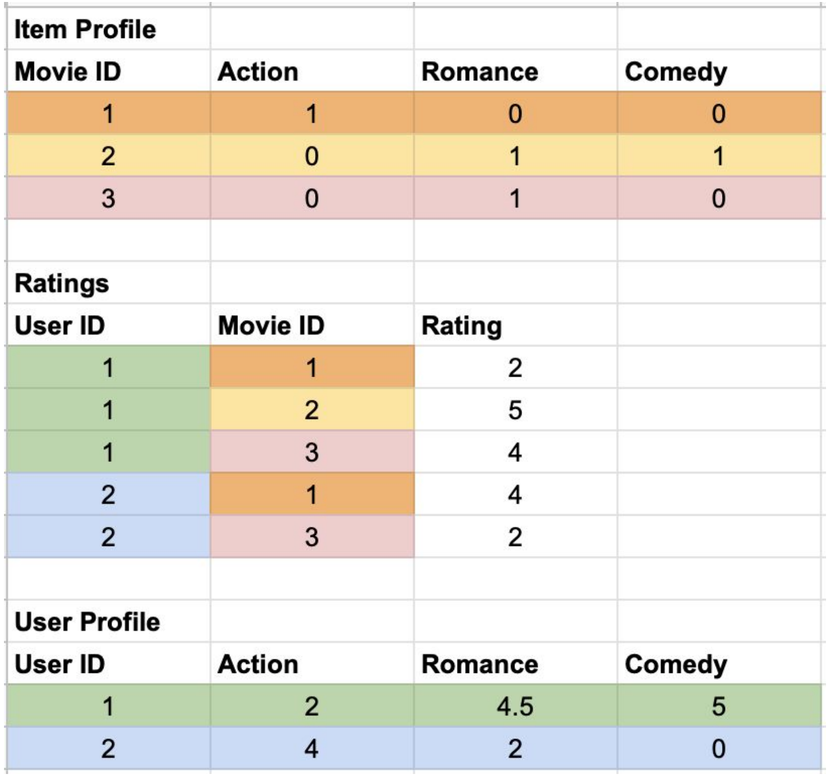
Content-Based Recommendation

1. User profile based on item profiles
    a. Genre
    b. Year of release of movie
2. Movie - Movie similarity

User and Item Vectors

Item Vector:
Vector of length total genres with 1 at relevant indices

User Vector:
Vector of length total genres with the value of average rating for each genre based on ratings in train set



Item Profile

Movie ID	Action	Romance	Comedy
1	1	0	0
2	0	1	1
3	0	1	0

Ratings

User ID	Movie ID	Rating
1	1	2
1	2	5
1	3	4
2	1	4
2	3	2

User Profile

User ID	Action	Romance	Comedy
1	2	4.5	5
2	4	2	0

Evaluation metrics

Metric	Content based (Genre)
RMSE	0.9185
MAE	0.7095

Metric	Content based (Genre)
Precision	0.800932214
Recall	0.495168862
F-Measure	0.6119842046
NDCG	0.945576877

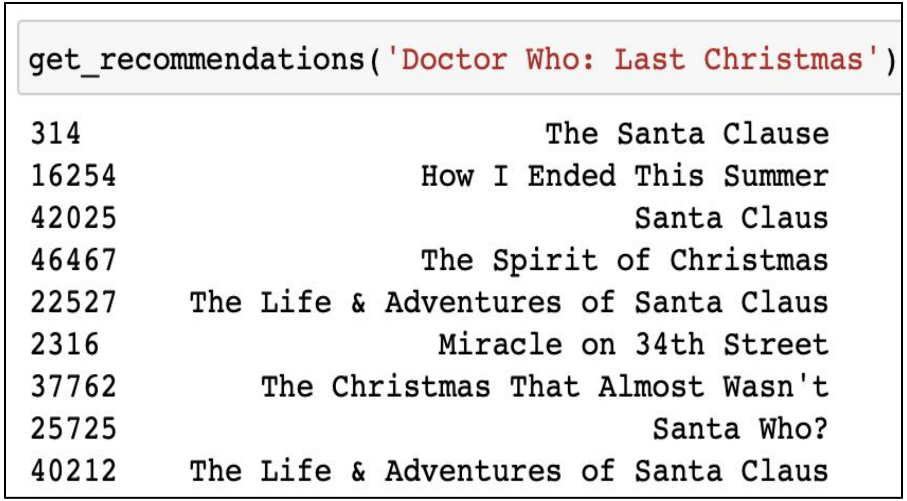
Movie-Movie Similarity

- TF-IDF using overview and tagline of movies (from TMDb)
- Issue: This just gives movies having similar description.

Movie-Movie Similarity (Cont.)

Overview of ‘Doctor Who: Last Christmas’

'The Doctor and Clara face their Last Christmas. Trapped on an Arctic base, under attack from terrifying creatures, who are you going to call? Santa Claus!'



```
get_recommendations('Doctor Who: Last Christmas')

314                              The Santa Clause
16254                        How I Ended This Summer
42025                                Santa Claus
46467                        The Spirit of Christmas
22527       The Life & Adventures of Santa Claus
2316                         Miracle on 34th Street
37762            The Christmas That Almost Wasn't
25725                                 Santa Who?
40212       The Life & Adventures of Santa Claus
```

Improvement

- Adding the genre two times to give more weightage
- Changing TF-IDF to Count Vector
  - TF-IDF gives lesser weight to frequently occurring terms across documents

Movie 1: '20 Years After'

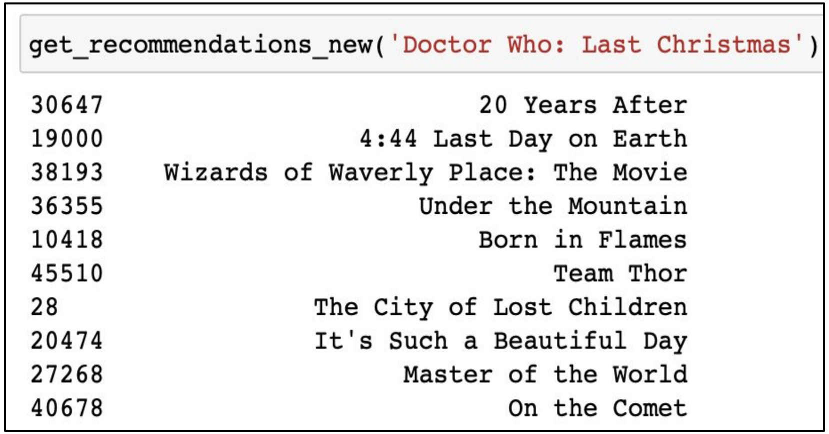
“In the middle of nowhere, 20 years after an apocalyptic terrorist event that obliterated the face of the world!”

Genre: ['Drama', 'Fantasy', 'Sci-Fi']

Movie 2: '4:44 Last Day on Earth'

Overview:
'A look at how a painter and a successful actor spend their last day together before the world comes to an end.'

Genre: ['Drama', 'Fantasy', 'Sci-Fi']



```
get_recommendations_new('Doctor Who: Last Christmas')

30647                             20 Years After
19000                     4:44 Last Day on Earth
38193        Wizards of Waverly Place: The Movie
36355                         Under the Mountain
10418                             Born in Flames
45510                                  Team Thor
28                       The City of Lost Children
20474                   It's Such a Beautiful Day
27268                        Master of the World
40678                               On the Comet
```

Doctor Who:

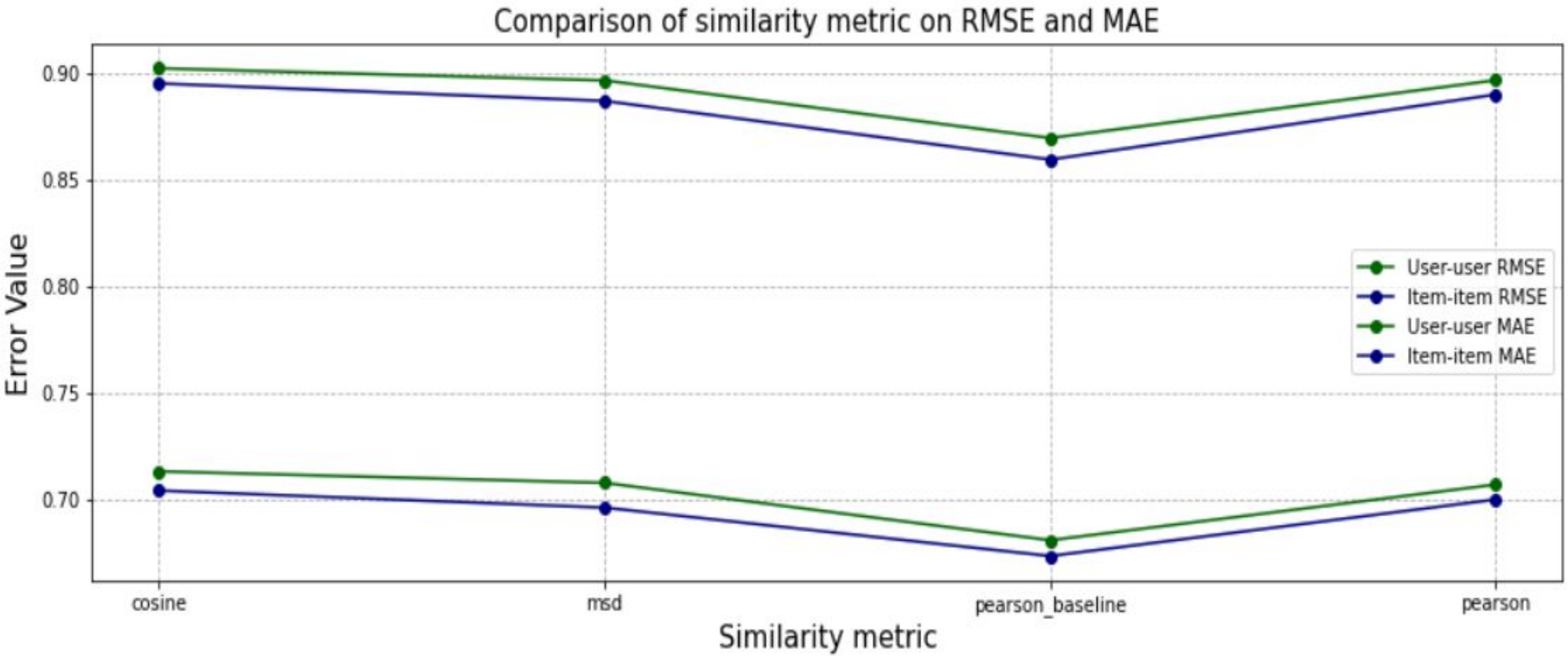
- 'The Doctor and Clara face their Last Christmas. Trapped on an Arctic base, under attack from terrifying creatures, who are you going to call? Santa Claus!'

- ['Adventure', 'Drama', 'Fantasy', 'Sci-Fi']

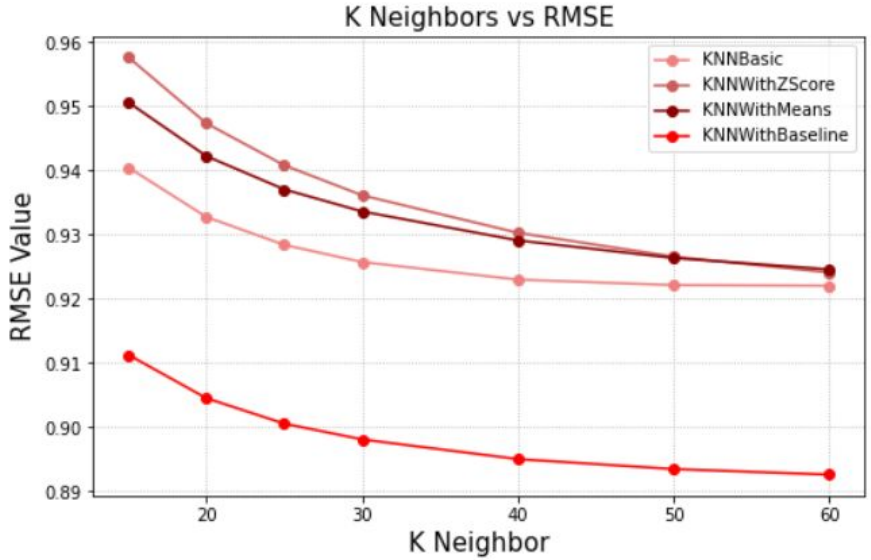
Collaborative Filtering

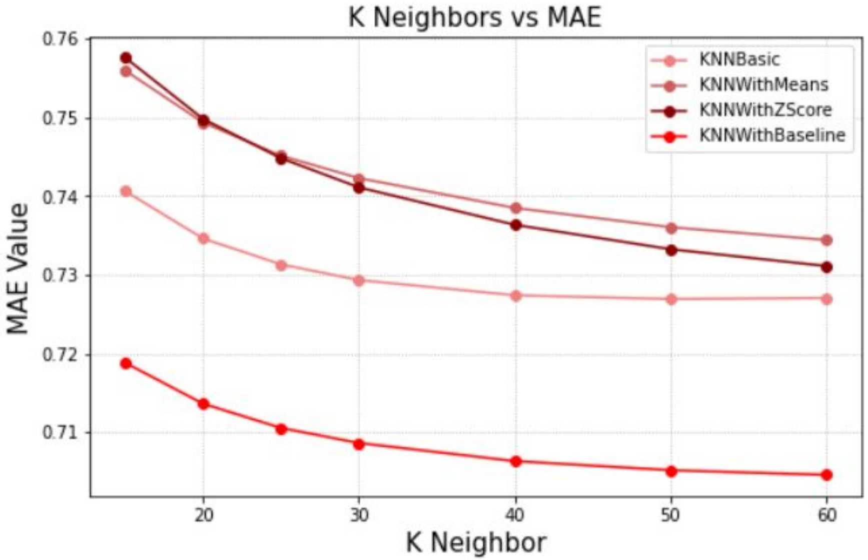
- KNN (k- nearest neighbors) algorithm using Surprise library
- Variations of KNN based approaches:
  - KNNBasic
  - KNNwithMeans
  - KNNWithZScore
  - KNNBaseline : integrates the baseline estimate ratings
- Similarity metrics:
  - Cosine similarity
  - Mean square difference based similarity
  - Pearson coefficient (mean-centered cosine similarity)
  - Pearson Baseline (uses global baselines for centering instead of means)

User-User and Item-Item comparison



RMSE vs MAE for different k

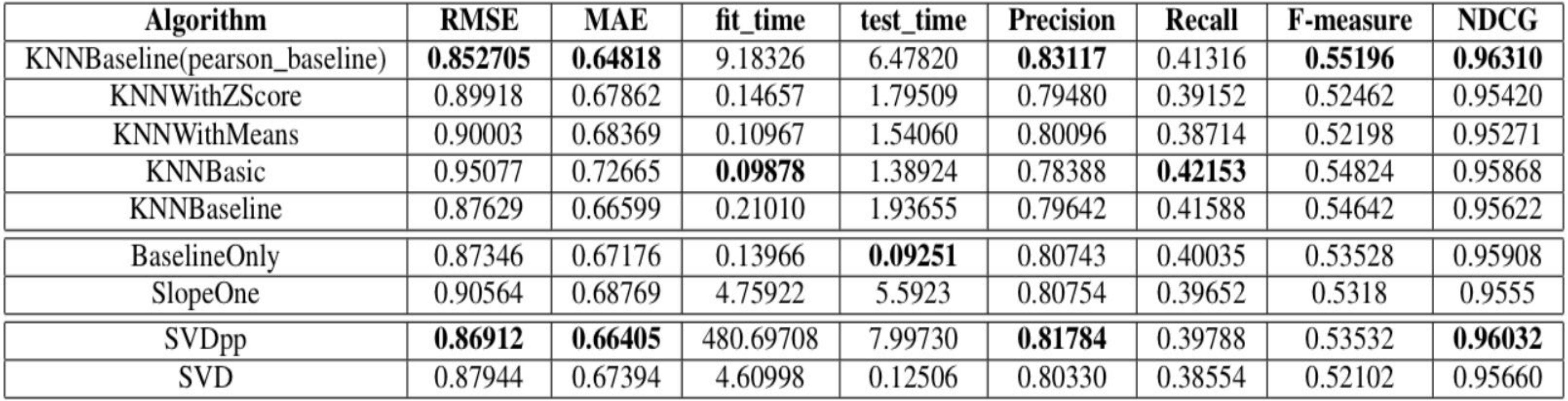




Latent Factor Methods

- Matrix Factorisation algorithms using Surprise library
  - SVD : baseline estimates + latent factor predictions
  - SVDpp : SVD + considers implicit ratings
- Hyperparameter tuning using GridsearchCV
  - Number of epochs, number of factors, regularization parameter

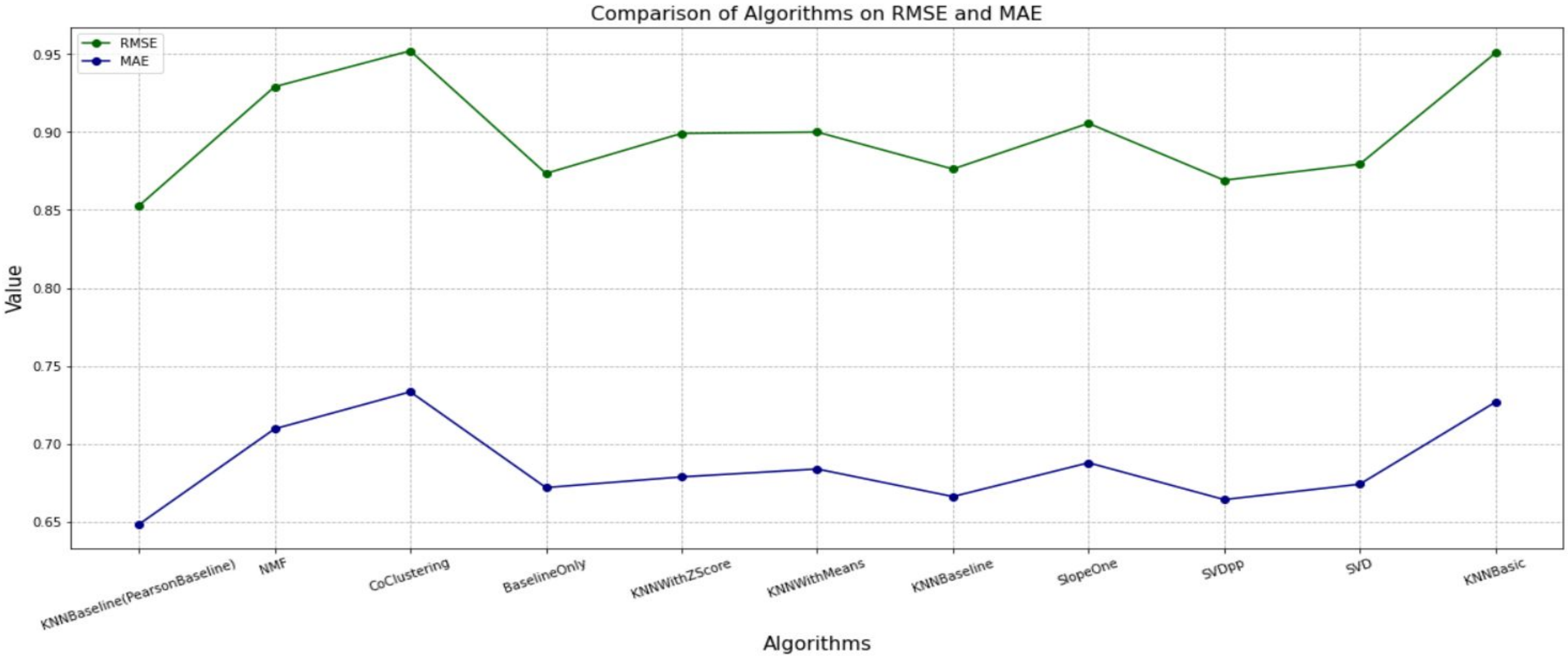
Evaluation of various algorithms:



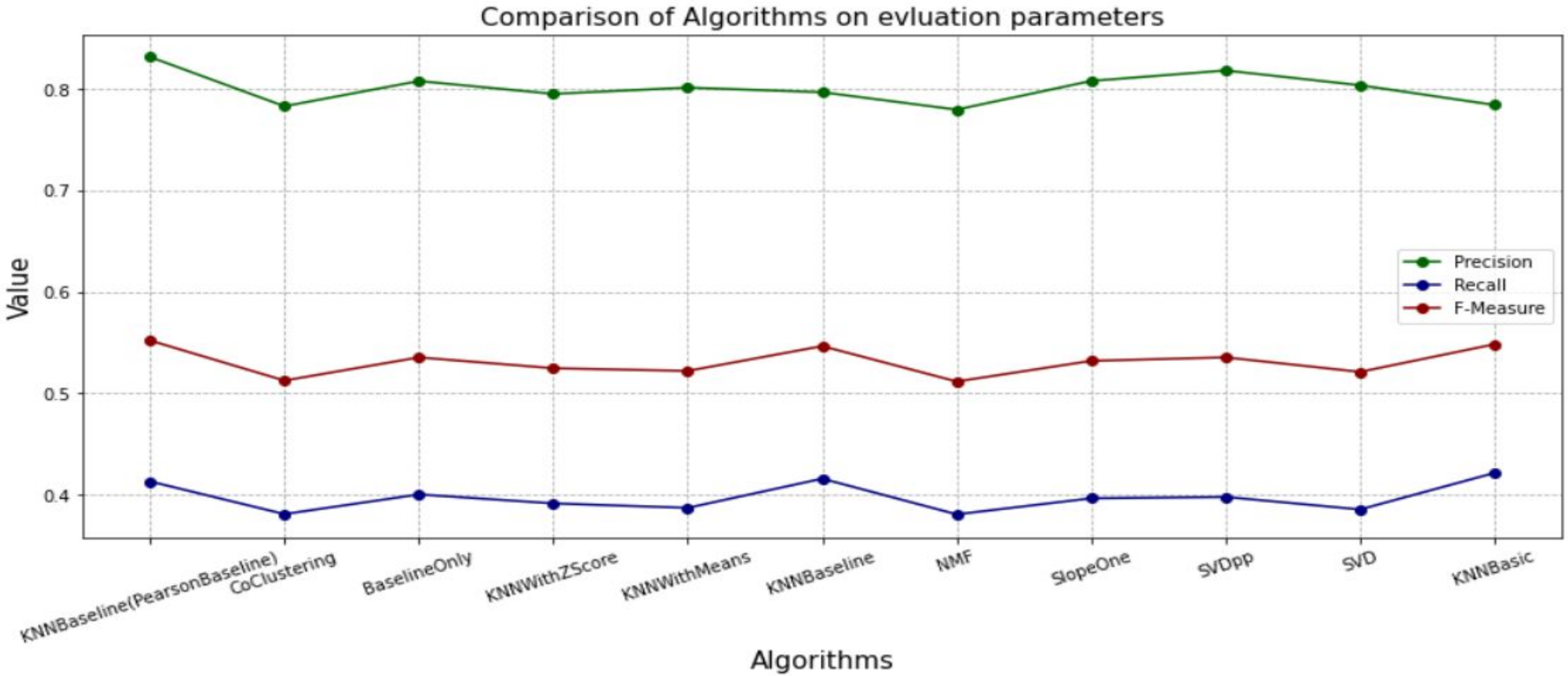
Algorithm	RMSE	MAE	fit_time	test_time	Precision	Recall	F-measure	NDCG
KNNBaseline(pearson_baseline)	**0.852705**	**0.64818**	9.18326	6.47820	**0.83117**	0.41316	**0.55196**	**0.96310**
KNNWithZScore	0.89918	0.67862	0.14657	1.79509	0.79480	0.39152	0.52462	0.95420
KNNWithMeans	0.90003	0.68369	0.10967	1.54060	0.80096	0.38714	0.52198	0.95271
KNNBasic	0.95077	0.72665	**0.09878**	1.38924	0.78388	**0.42153**	0.54824	0.95868
KNNBaseline	0.87629	0.66599	0.21010	1.93655	0.79642	0.41588	0.54642	0.95622
BaselineOnly	0.87346	0.67176	0.13966	**0.09251**	0.80743	0.40035	0.53528	0.95908
SlopeOne	0.90564	0.68769	4.75922	5.5923	0.80754	0.39652	0.5318	0.9555
SVDpp	**0.86912**	**0.66405**	480.69708	7.99730	**0.81784**	0.39788	0.53532	**0.96032**
SVD	0.87944	0.67394	4.60998	0.12506	0.80330	0.38554	0.52102	0.95660

Precision and Recall @ 5
Relevant : rating >=3.75

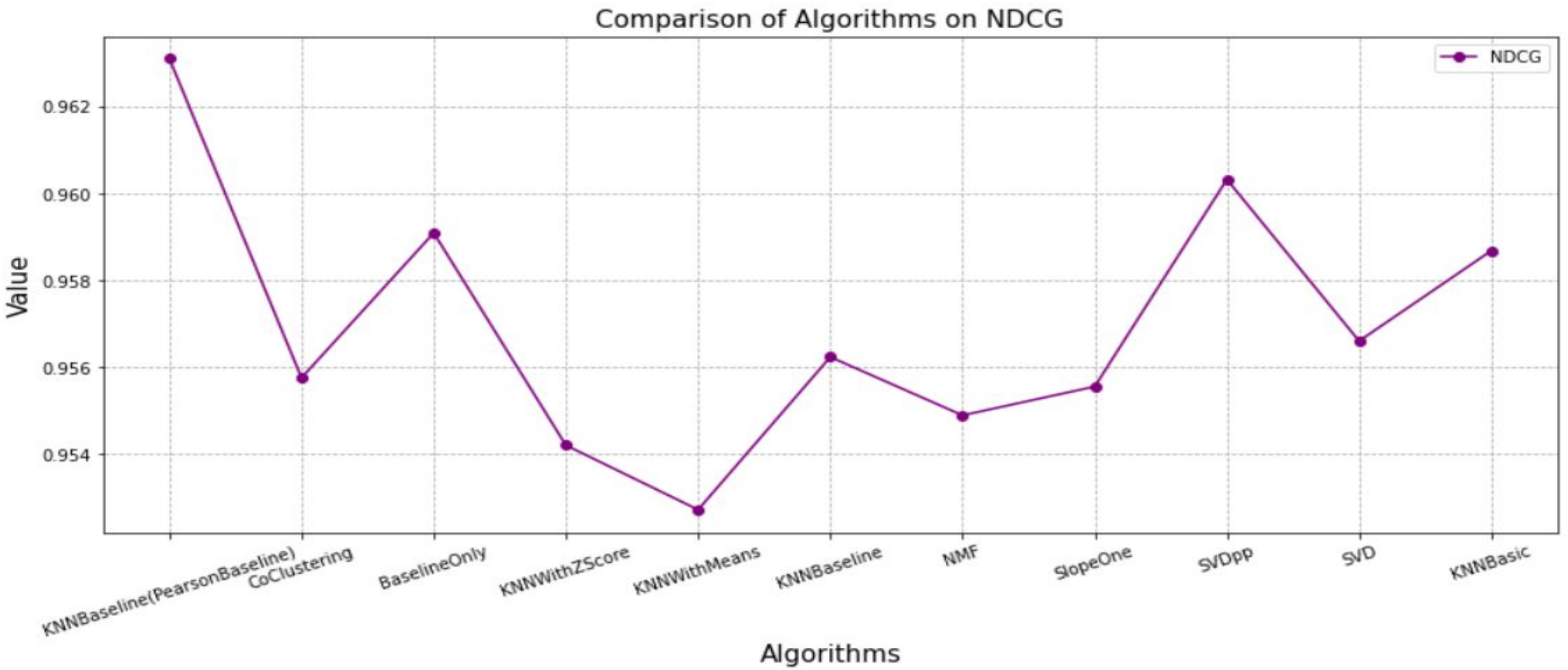
RMSE vs MAE for different algorithms



Evaluation of different algorithms

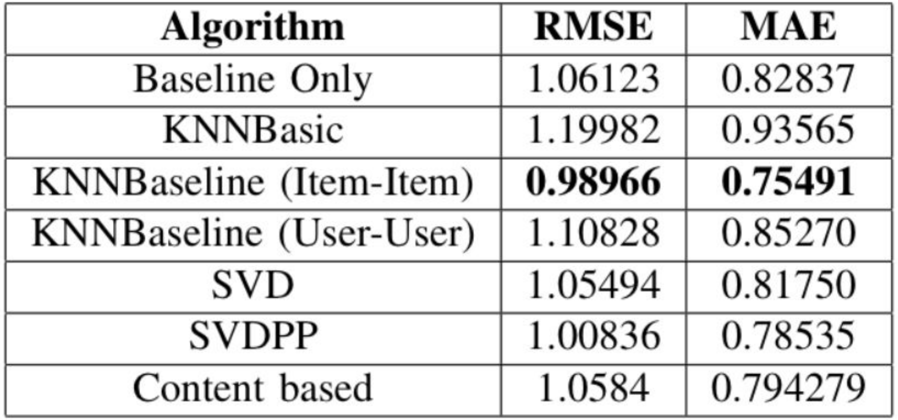


NDCG scores for different algorithms



Which model is best for less ratings in training data?

(Less than 18 ratings per user)



Algorithm	RMSE	MAE
Baseline Only	1.06123	0.82837
KNNBasic	1.19982	0.93565
KNNBaseline (Item-Item)	**0.98966**	**0.75491**
KNNBaseline (User-User)	1.10828	0.85270
SVD	1.05494	0.81750
SVDPP	1.00836	0.78535
Content based	1.0584	0.794279

Combined Model

- Matrix Factorization + CF
- Weighted linear combination of prediction ratings
- Combined:
  - KNNBaseline (with pearson baseline similarity)
  - SVDpp
  - SVD
  - BaselineOnly

Weighted Combination of the best Algorithms

SVD	KNNBaseline (PearsonBaseline)	SVDPP	Baseline	RMSE	MAE	Precision	Recall	F-Measure	NDCG
0	0	1	0	0.8626	0.6735	0.79921	0.33642	0.47352	0.95872
0	1	0	0	0.8527	0.6482	0.81984	0.36242	0.50264	0.96310
0	0.6	0.4	0	0.844	0.6427	0.82732	0.35428	0.49611	0.96637
0.05	0.6	0.35	0	0.84405	0.6427	0.82836	0.35424	0.49626	0.96731
0.1	0.5	0.3	0.1	0.84405	0.6433	0.83139	0.35040	0.49301	0.96614

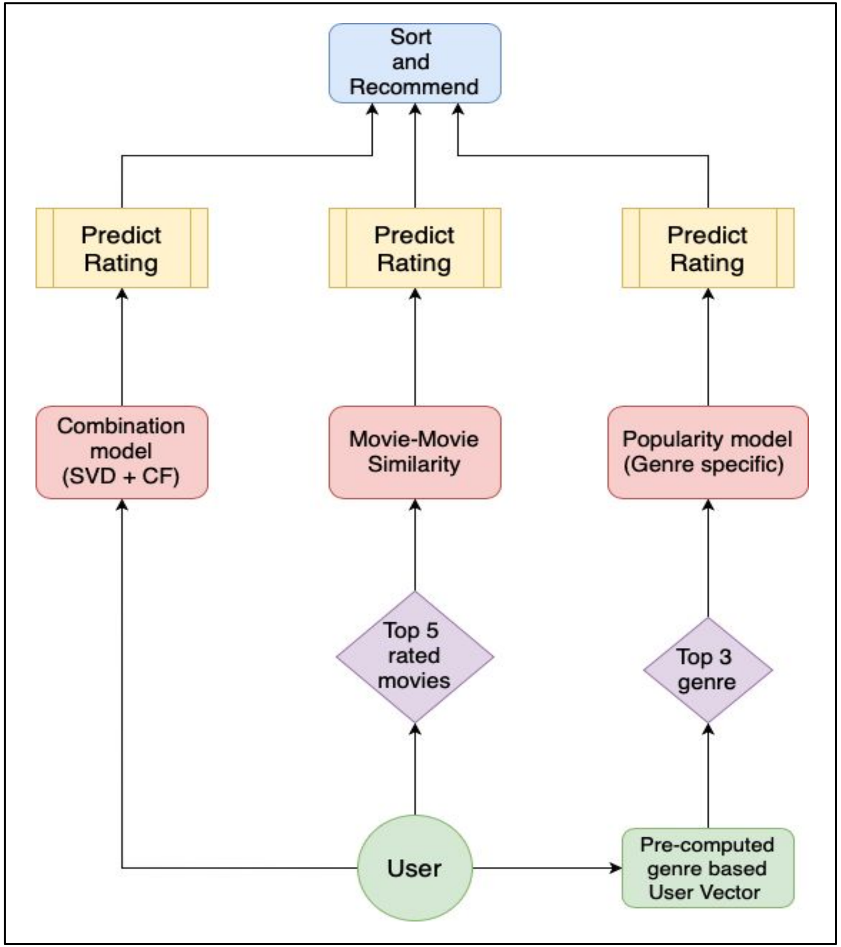
Hybrid Model

Combination of recommendations using:

1. Combined model (SVD + CF)
2. Content Based Movie-Movie Similarity
3. Popularity model

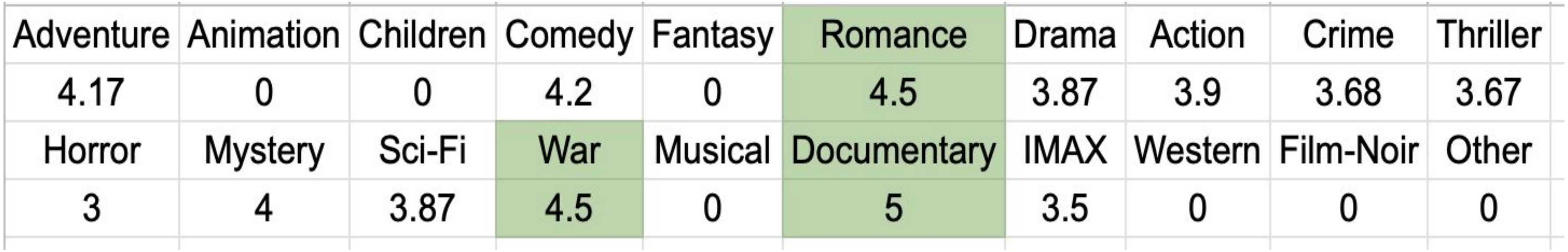
   +

   User Profile (Genre Based)

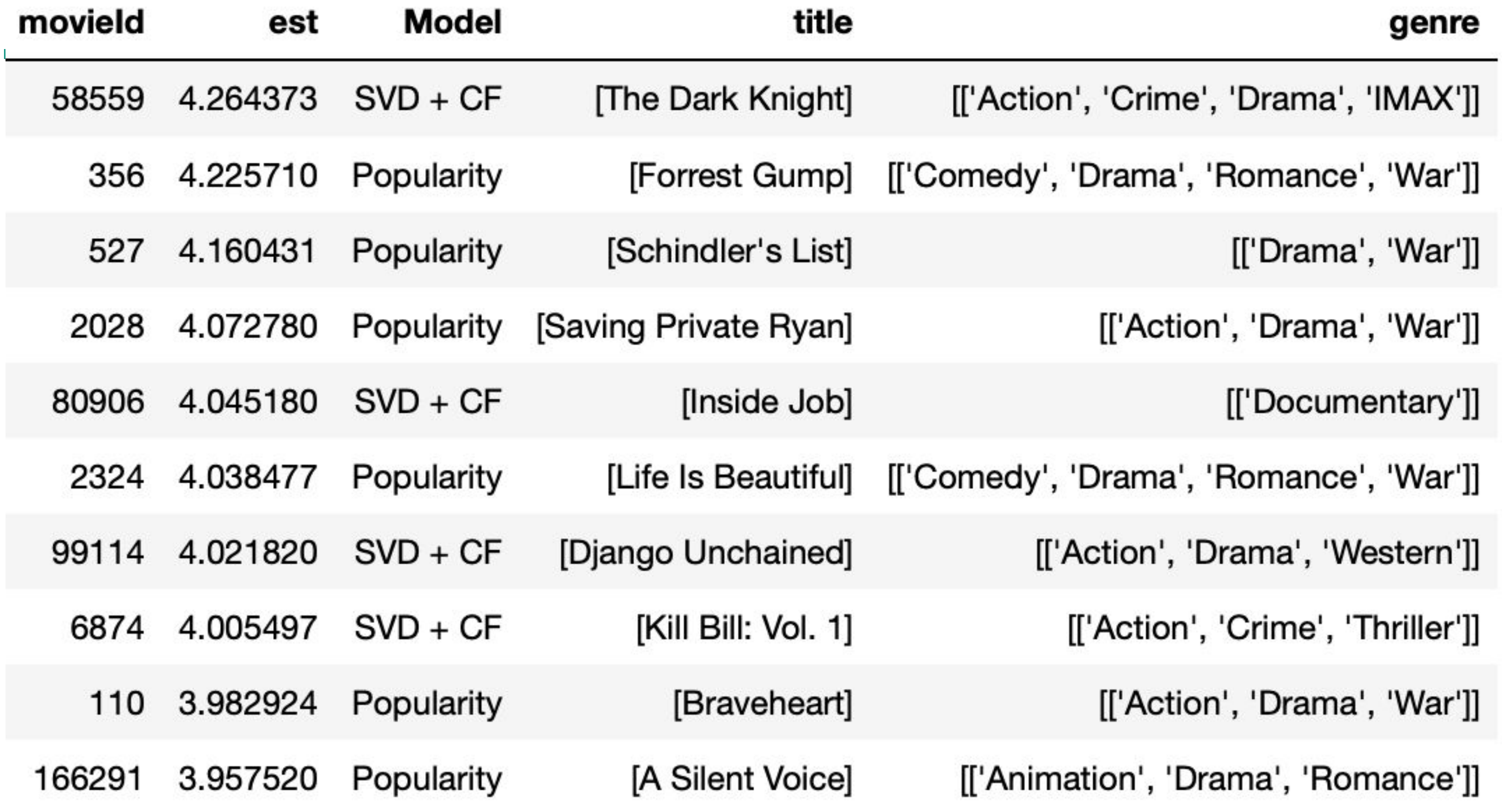


Results when enough information of User

- User Id = 2
- User top genre list from User vector:
    - ['Documentary', 'Romance', 'War']:



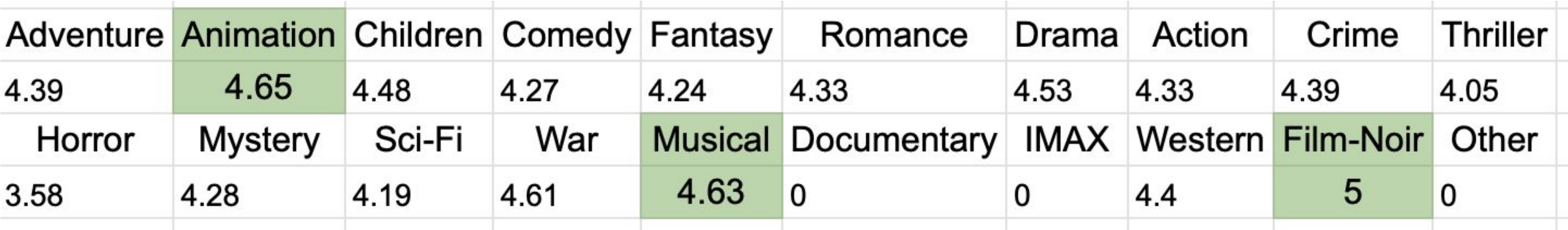
Adventure	Animation	Children	Comedy	Fantasy	Romance	Drama	Action	Crime	Thriller
4.17	0	0	4.2	0	4.5	3.87	3.9	3.68	3.67
Horror	Mystery	Sci-Fi	War	Musical	Documentary	IMAX	Western	Film-Noir	Other
3	4	3.87	4.5	0	5	3.5	0	0	0



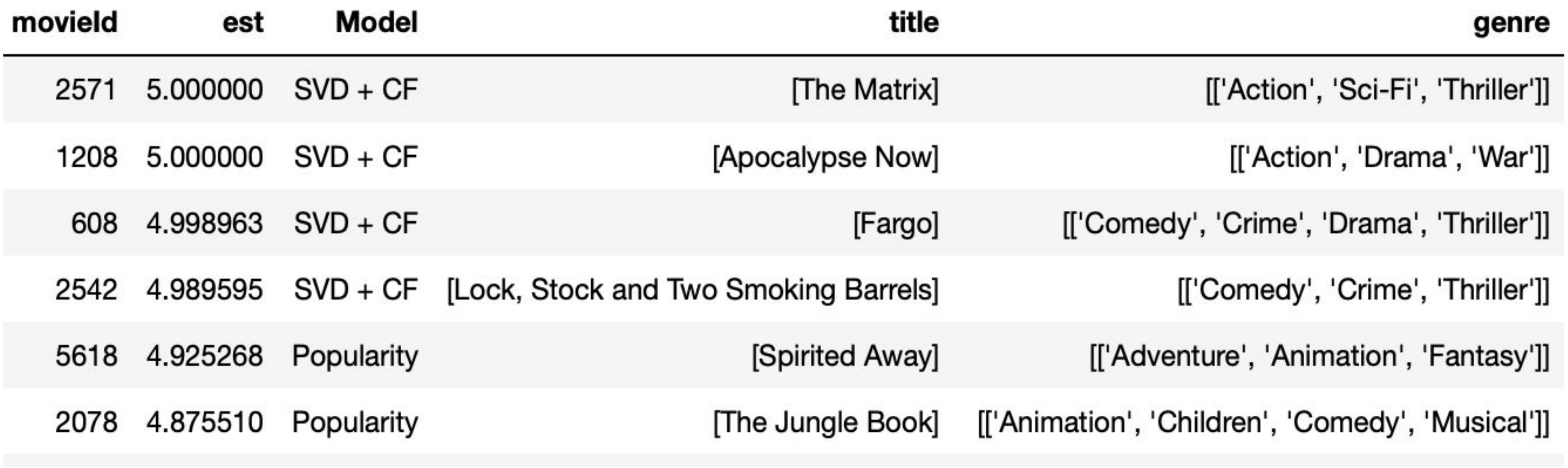
movieId	est	Model	title	genre
58559	4.264373	SVD + CF	[The Dark Knight]	[['Action', 'Crime', 'Drama', 'IMAX']]
356	4.225710	Popularity	[Forrest Gump]	[['Comedy', 'Drama', 'Romance', 'War']]
527	4.160431	Popularity	[Schindler's List]	[['Drama', 'War']]
2028	4.072780	Popularity	[Saving Private Ryan]	[['Action', 'Drama', 'War']]
80906	4.045180	SVD + CF	[Inside Job]	[['Documentary']]
2324	4.038477	Popularity	[Life Is Beautiful]	[['Comedy', 'Drama', 'Romance', 'War']]
99114	4.021820	SVD + CF	[Django Unchained]	[['Action', 'Drama', 'Western']]
6874	4.005497	SVD + CF	[Kill Bill: Vol. 1]	[['Action', 'Crime', 'Thriller']]
110	3.982924	Popularity	[Braveheart]	[['Action', 'Drama', 'War']]
166291	3.957520	Popularity	[A Silent Voice]	[['Animation', 'Drama', 'Romance']]

Results when enough information of User

- User Id = 1
- User top genre list from User vector:
    - [‘Film-Noir’, ‘Animation’, ‘Musical’]:



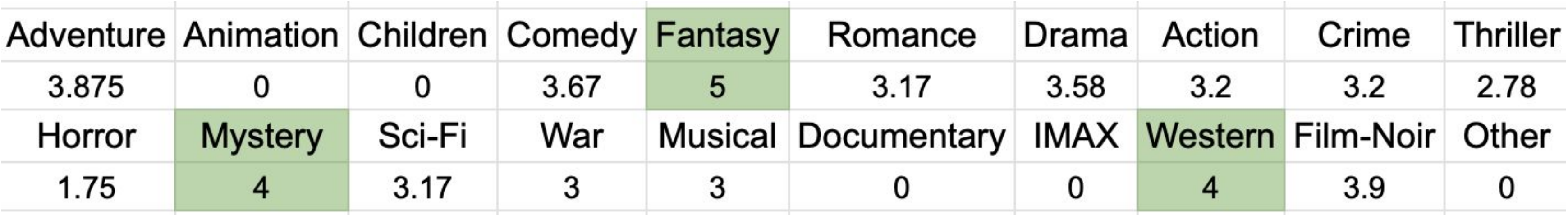
Adventure	Animation	Children	Comedy	Fantasy	Romance	Drama	Action	Crime	Thriller
4.39	4.65	4.48	4.27	4.24	4.33	4.53	4.33	4.39	4.05
Horror	Mystery	Sci-Fi	War	Musical	Documentary	IMAX	Western	Film-Noir	Other
3.58	4.28	4.19	4.61	4.63	0	0	4.4	5	0



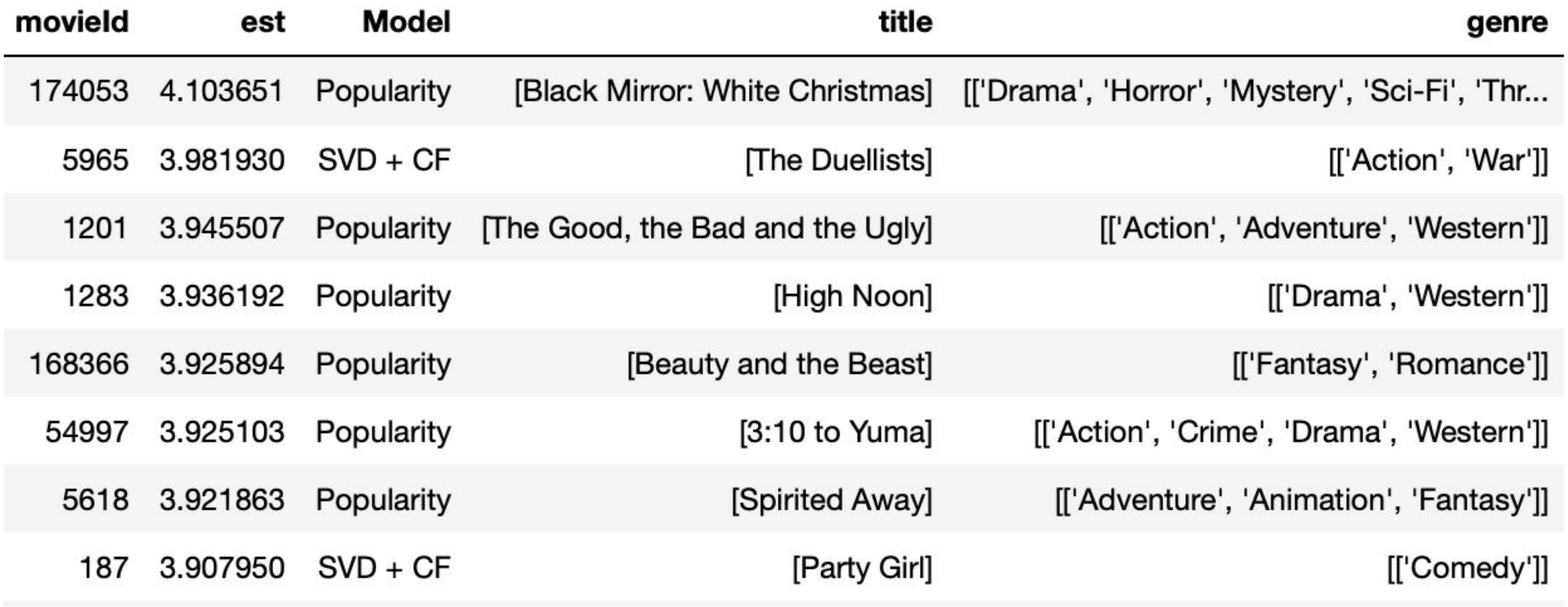
movieId	est	Model	title	genre
2571	5.000000	SVD + CF	[The Matrix]	[['Action', 'Sci-Fi', 'Thriller']]
1208	5.000000	SVD + CF	[Apocalypse Now]	[['Action', 'Drama', 'War']]
608	4.998963	SVD + CF	[Fargo]	[['Comedy', 'Crime', 'Drama', 'Thriller']]
2542	4.989595	SVD + CF	[Lock, Stock and Two Smoking Barrels]	[['Comedy', 'Crime', 'Thriller']]
5618	4.925268	Popularity	[Spirited Away]	[['Adventure', 'Animation', 'Fantasy']]
2078	4.875510	Popularity	[The Jungle Book]	[['Animation', 'Children', 'Comedy', 'Musical']]

Results when less information about user

- User Id = 9
- User top genre from User Vector:
  - [‘Fantasy’, ‘Western’, ‘Mystery’]



Adventure	Animation	Children	Comedy	Fantasy	Romance	Drama	Action	Crime	Thriller
3.875	0	0	3.67	5	3.17	3.58	3.2	3.2	2.78
Horror	Mystery	Sci-Fi	War	Musical	Documentary	IMAX	Western	Film-Noir	Other
1.75	4	3.17	3	3	0	0	4	3.9	0



movieId	est	Model	title	genre
174053	4.103651	Popularity	[Black Mirror: White Christmas]	[['Drama', 'Horror', 'Mystery', 'Sci-Fi', 'Thr...
5965	3.981930	SVD + CF	[The Duellists]	[['Action', 'War']]
1201	3.945507	Popularity	[The Good, the Bad and the Ugly]	[['Action', 'Adventure', 'Western']]
1283	3.936192	Popularity	[High Noon]	[['Drama', 'Western']]
168366	3.925894	Popularity	[Beauty and the Beast]	[['Fantasy', 'Romance']]
54997	3.925103	Popularity	[3:10 to Yuma]	[['Action', 'Crime', 'Drama', 'Western']]
5618	3.921863	Popularity	[Spirited Away]	[['Adventure', 'Animation', 'Fantasy']]
187	3.907950	SVD + CF	[Party Girl]	[['Comedy']]

Takeaways:

1. Content based with genre is good when a user has less ratings.
2. Movie similarity metric based on features like overview, taglines and genre.
3. Item-item collaborative filtering works better than user-user collaborative filtering.
4. KNN based and SVD algorithms improve when global baselines are added.
5. Combining the predictions and recommendations of different models gives better results in terms of accuracy and quality of recommendations.

References:

1. Surprise library: https://surprise.readthedocs.io/en/stable/
2. Hybrid recommendation system: https://arxiv.org/pdf/1901.03888.pdf
3. Evaluating recommendation system: http://fastml.com/evaluating-recommender-systems/

Thank you!

Questions?