OPTIMIZING MOVIE RECOMMENDATION FOR ENHANCED USER EXPERIENCE.

Ganya Janardhan (A20517083)

Jayanth Chidananda (A20517012)

Vismaya M (A20519405)

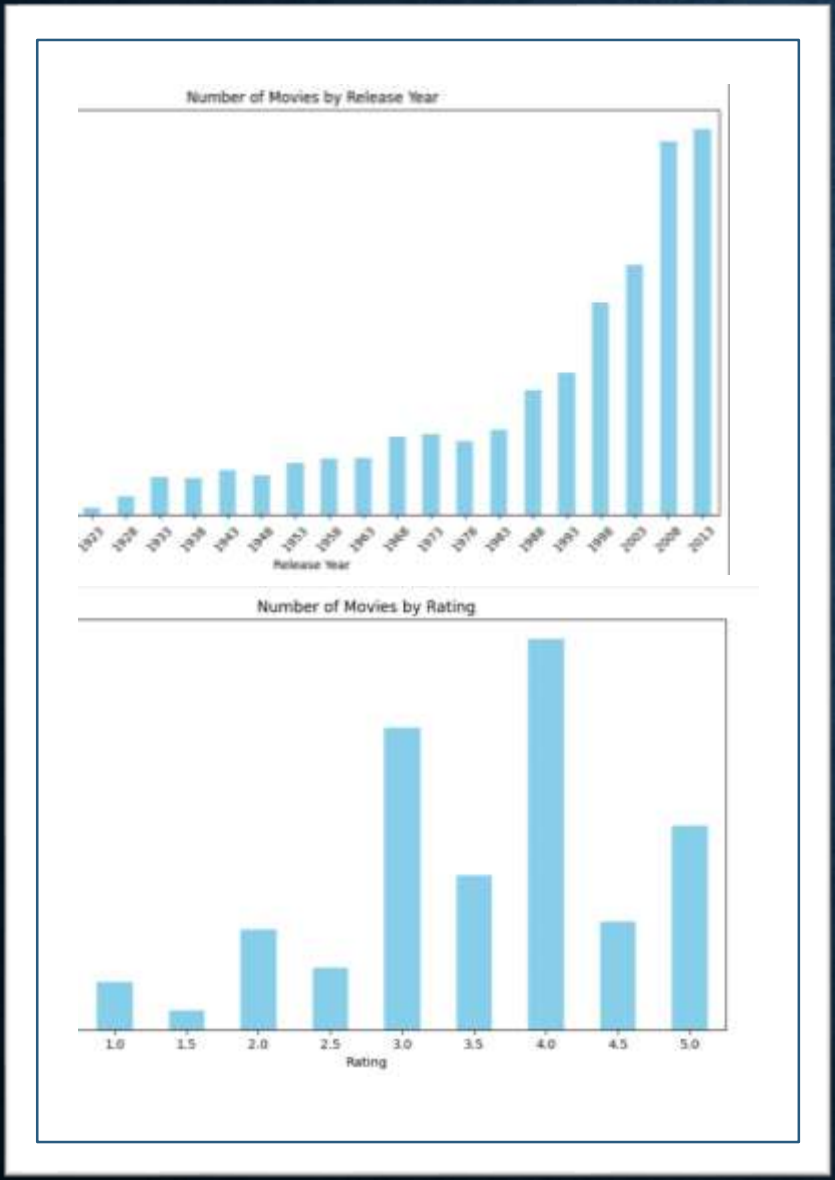
PROBLEM STATEMENT

- **Goal:** Develop a recommendation system utilizing the MovieLens (20M) dataset to offer movie suggestions derived from users' historical interactions.
- Numerous e-commerce platforms leverage recommendation systems to offer personalized recommendations to their users.
- These systems analyze users' browsing history and preferences to deliver tailored suggestions.
- In this project we will use user-user based collaborative filtering technique (Matrix Factorization and Cosine similarity) and Item-item based collaborative filtering (KNN) to predict recommendations to users.

DATA SET

- Movielens (20M) dataset from Kaggle was used. However, due to local RAM restrictions, only 4M data points were used.

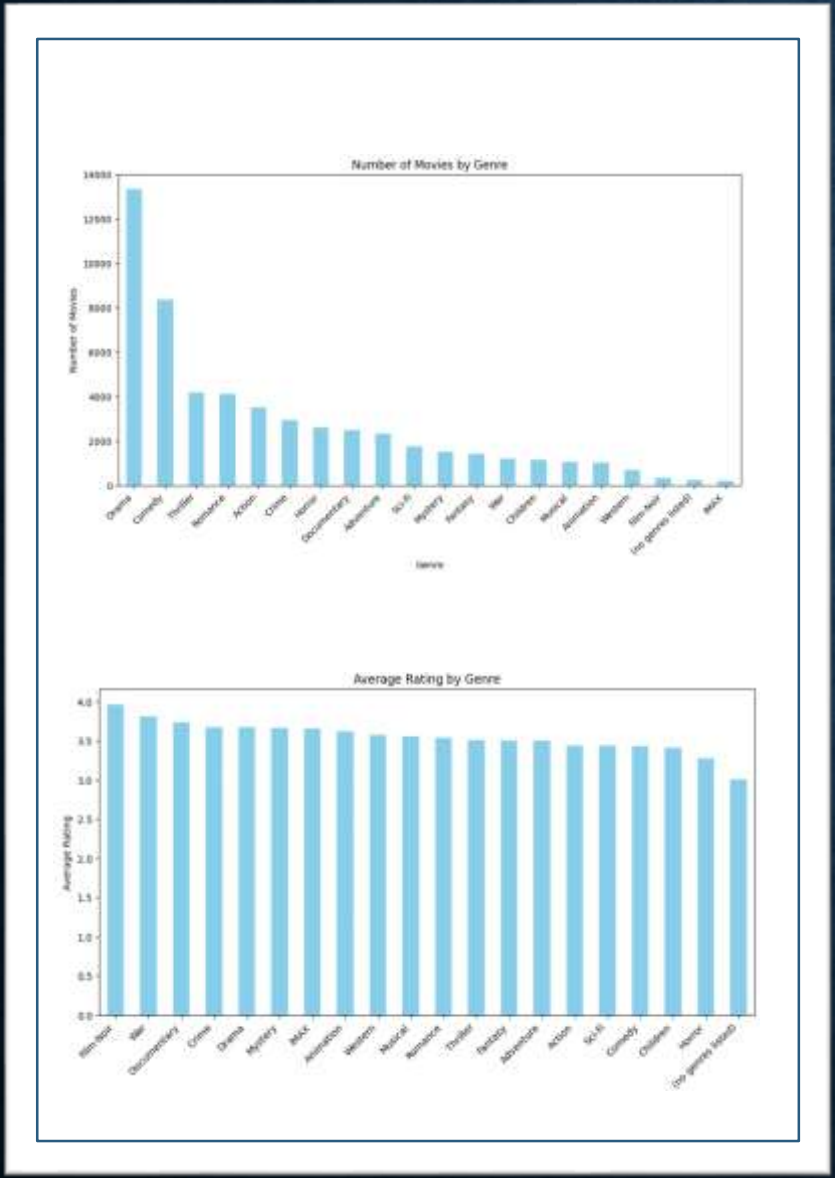
File	*Size (n×p)*	*Features*	*Description*
Ratings	20M*4	'userId', 'movieId', 'Rating', 'timestamp'	Contains User IDs and ratings for movies. Ratings are provided on 5-star scale
Movie	27K*3	'movieId', 'title', 'genres'	It has a genre information and also used as lookup to identify movies
Tags	465K*4	'userId', 'movieId','tag', 'timestamp'	These are user generated tags for movies
Genome Tags	1128*2	'tagId', 'tag';	Tga description are provided for TagIDs
Genome Scores	11M*3	'movieId','tagId', 'relevance',	Relevance score is provided for a tag with associated movie
Link		'movieId', imdbId', 'tmdbid'	This dataset contains identifiers for linking to other sources



DATA VISUALIZATION

- The figure above depicts the number of movies in the dataset each year; as shown in the plot, the number of movies in the 2000s is substantially higher than in prior years.

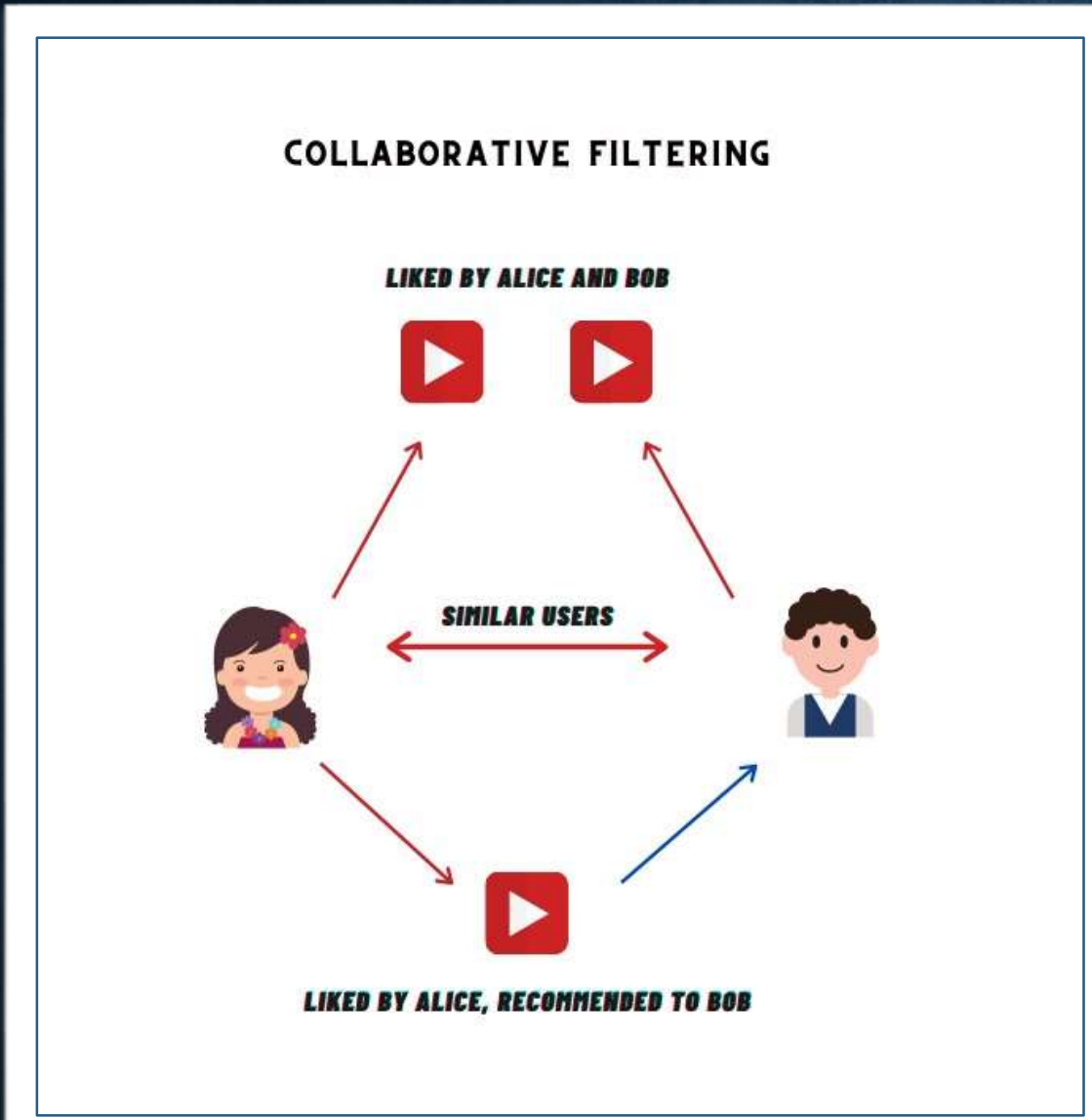
- The plot above gives a visualization of the number of movies and their ratings. Maximum number of movies have a rating of 4, followed by 3



DATA VISUALIZATION

- The plot above visualizes the movies and their genres. It is noticeable that the maximum number of movies are classified as “Drama”.

- The plot above depicts the average rating of all movies by genre; it can be seen that the average rating in each genre ranges from 3 to 4.



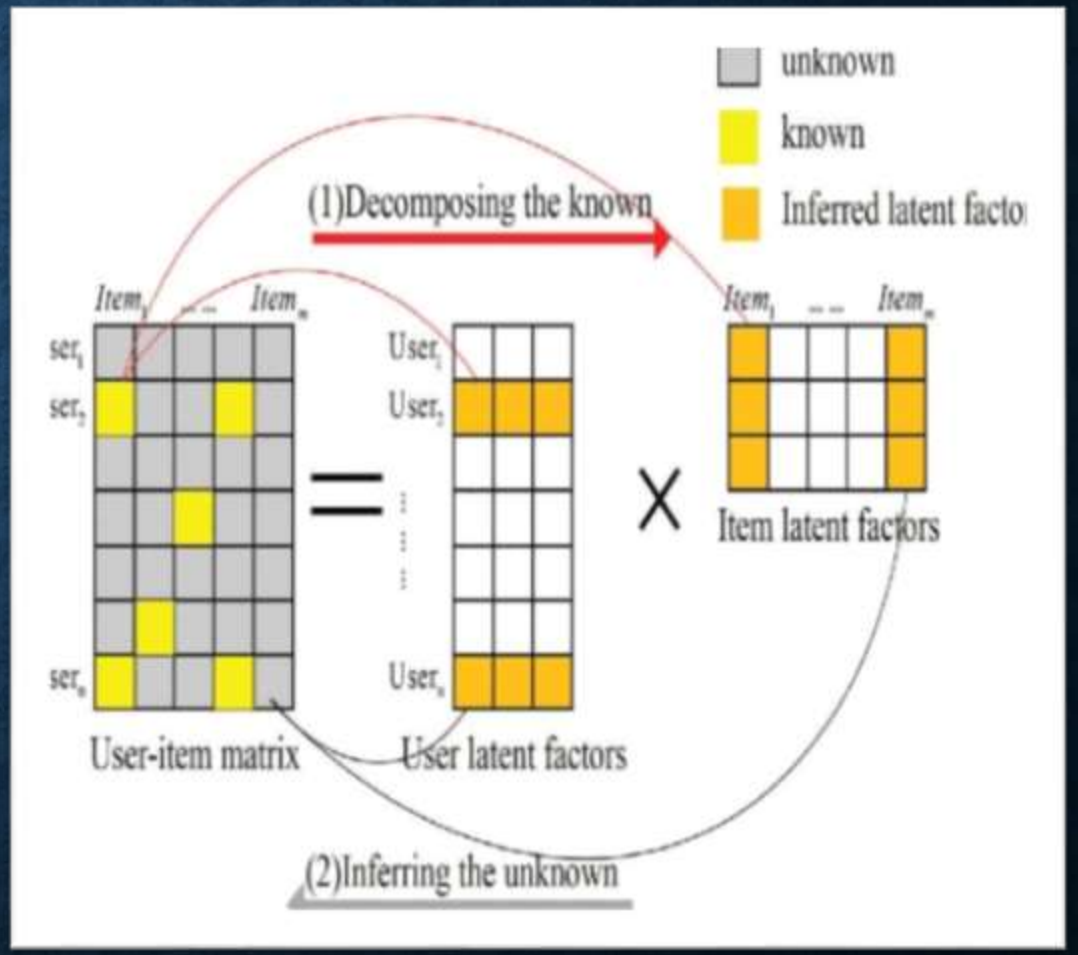
METHODOLOGY

COLLABORATIVE FILTERING:

- User A has provided high ratings for movies M1 and M2 after watching them.
- User B has also watched movie M2 and has given it a favorable rating.
- Subsequently, the collaborative filtering model recommends movie M1 to User B based on the shared preferences with User A.

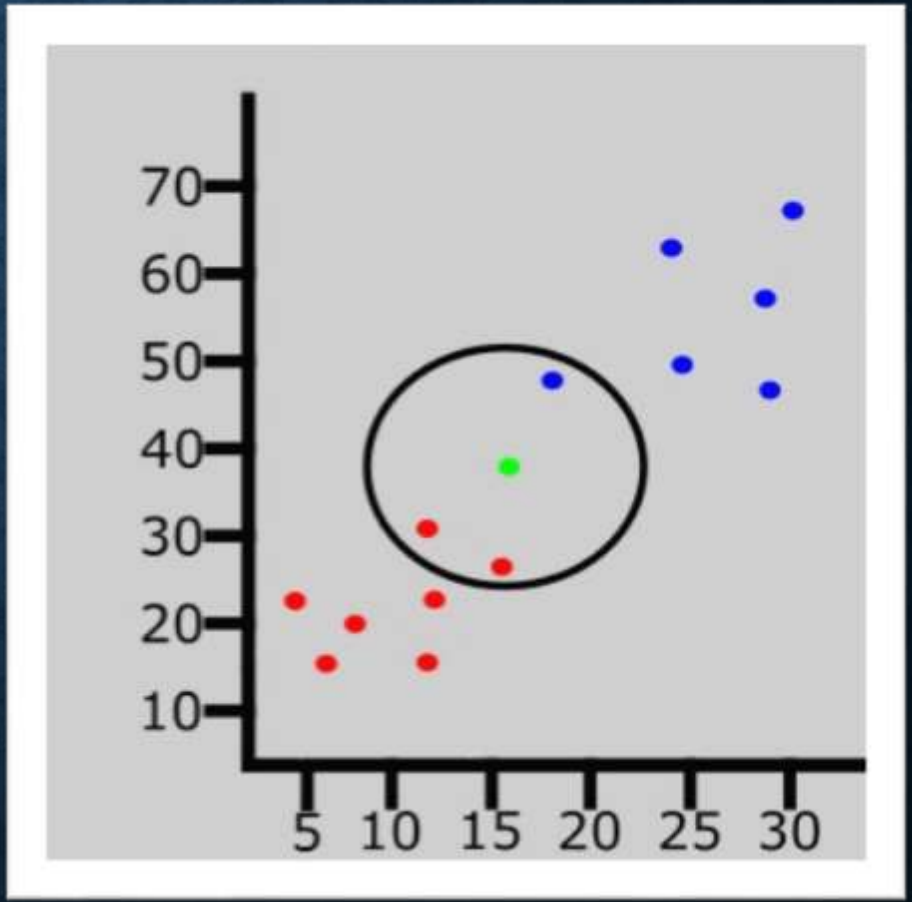
MATRIX FACTORIZATION

- Generate User-movie matrix
- Identify the optimal number of latent features for the recommendation system.
- Split the rating matrix into two matrices: one for user-related features (P) and the other for movie-related features (Q).
- Utilize gradient descent to iteratively determine the matrices P and Q, optimizing the given equation.



K-NEAREST NEIGHBOURS

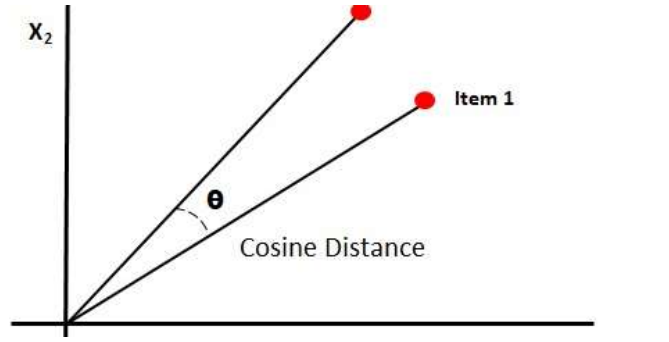
- Implemented item-based collaborative filtering to recommend movies based on user ratings and similarities between items.
- Conducted a model-building stage to calculate similarities between all pairs of movies, employing metrics like correlation or cosine similarity
- Executed a recommendation stage using the most similar movies to a user's preferences, generating suggestions through a weighted sum or linear regression for personalized recommendations.



COSINE SIMILARITY

- Assess the similarity between two users by comparing the angle (θ) between the vectors representing them.
- A smaller θ indicates a lower angle between the user vectors, suggesting a higher degree of similarity between the users.

$$\vartheta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} E}}$$



METRICS FOR EVALUATION

Matrix factorization :

For matrix factorization, Mean Absolute Error (MAE) and Mean Cubic Error (MCE) were used to compare the train and test sets. Mean Absolute Error would give an idea of the absolute error on our predictions and Mean Cubic Error would help us penalize predictions that are far off from the true value. Comparing both metrics would give an idea on the performance of the model.

RESULTS:

The results provided by collaborative filtering (method 1), cosine similarity (method 2) and KNN (method 3) are similar. Let us consider an example to clearly explain this. Here, the user has watched the following movies, and based on this, we make the following observations about both methods.

Sample 1

User ID= 123 has rated the following movies

movieId <chr>	rating <dbl>	title <chr>	genres <chr>
778	5	Trainspotting (1996)	Comedy|Crime|Drama
10	3	GoldenEye (1995)	Action|Adventure|Thriller
104	3	Happy Gilmore (1996)	Comedy
1103	3	Rebel Without a Cause (1955)	Drama
21	3	Get Shorty (1995)	Comedy|Crime|Thriller
762	3	Striptease (1996)	Comedy|Crime
135	1	Down Periscope (1996)	Comedy
736	1	Twister (1996)	Action|Adventure|Romance|Thriller
849	1	Escape from L.A. (1996)	Action|Adventure|Sci-Fi|Thriller

Using Matrix factorization, the model recommends the following top movies:

Movies recommended by Matrix Factorization

movieId <chr>	rating <dbl>	title <chr>
1089	3.90	Reservoir Dogs (1992)
1094	3.94	Crying Game, The (1992)
1136	3.91	Monty Python and the Holy Grail (1975)
1206	4.03	Clockwork Orange, A (1971)
1223	3.85	Grand Day Out with Wallace and Gromit, A (1989)
1272	3.96	Patton (1970)
1298	4.13	Pink Floyd: The Wall (1982)
1729	3.87	Jackie Brown (1997)
194	3.87	Smoke (1995)
2076	3.93	Blue Velvet (1986)

1-10 of 30 rows | 1-3 of 4 columns

Using Cosine Similarity, the model recommends the following top movies:

Movies recommended by Cosine Similarity

movieId <chr>	rating <dbl>	title <chr>	genres <chr>
1080	4	Monty Python's Life of Brian (1979)	Comedy
1185	5	My Left Foot (1989)	Drama
1193	5	One Flew Over the Cuckoo's Nest (1975)	Drama
1208	5	Apocalypse Now (1979)	Action|Drama|War
1212	5	Third Man, The (1949)	Film-Noir|Mystery|Thriller
1220	5	Blues Brothers, The (1980)	Action|Comedy|Musical
1231	5	Right Stuff, The (1983)	Drama
1249	4	Femme Nikita, La (Nikita) (1990)	Action|Crime|Romance|Thriller
1261	5	Evil Dead II (Dead by Dawn) (1987)	Action|Comedy|Fantasy|Horror
1262	4	Great Escape, The (1963)	Action|Adventure|Drama|War

Using KNN, the model recommends the following top movies:

Movies recommended KNN

```
First $20 Million Is Always the Hardest, The
Bottle Rocket
Ghost and the Darkness, The
Lt. Robin Crusoe, U.S.N.
Browning Version, The
Faat Kiné
Die Hard 2
Women Without Men (Zanan-e bedun-e mardan)
Why Worry?
```

Both the models Matrix Factorization and Cosine Similarity recommended a movie from the series Monty Python.
Meanwhile KNN suggested different movies from the Comedy genre.

CONCLUSION

- Each of the three models offers commendable movie recommendation by leveraging the user's viewing history.
- Three models have the capability to predict any desired number of movie recommendations.
- In terms of user recommendations, the tested models appear to provide similar suggestions, particularly within the same genre.
- However, it's worth noting that Matrix factorization demands more memory and time for execution compared to Cosine Similarity and it is susceptible to sparse data when compared to KNN.

FUTURE WORK

- In future work, an enhanced movie recommendation system could be developed by combining the strengths of KNN, matrix factorization, and cosine similarity.
- Integration of these methods could involve leveraging KNN for its simplicity and interpretability, matrix factorization for handling sparse data and scalability, and cosine similarity for robustness to varying scales.
- Explore content-based filtering and hybrid recommenders for enhanced recommendation strategies.

THANK YOU