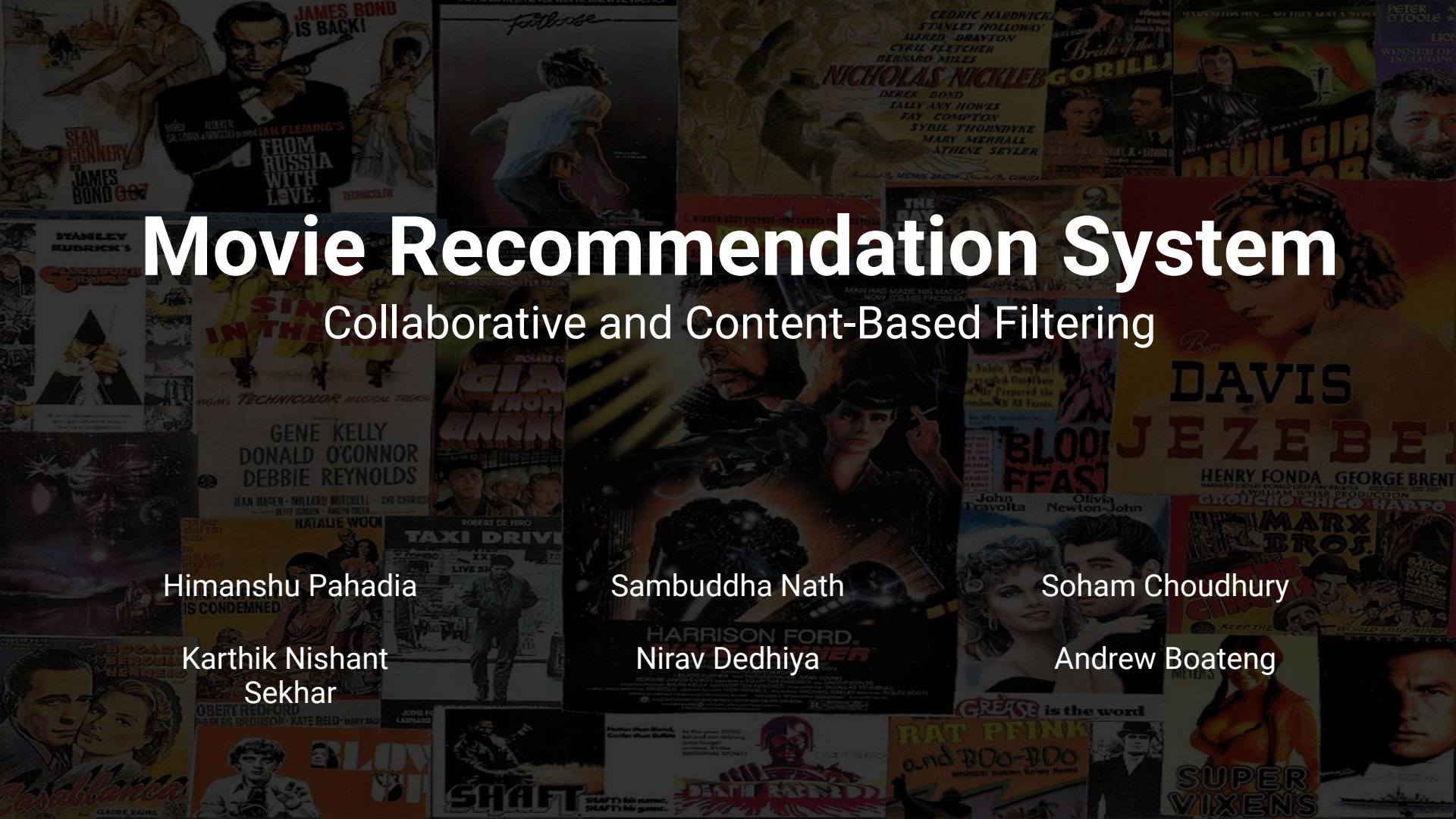


Movie Recommendation System

Collaborative and Content-Based Filtering

Himanshu Pahadia

Sambuddha Nath

Soham Choudhury

Karthik Nishant Sekhar

Nirav Dedhiya

Andrew Boateng

Introduction

Problem Summary

- Build a recommendation system to provide the best movie choices for a user based on their taste.
- This has in recent times become a very important competitive tool used by websites which have a wide range of products and services to offer its consumers.
- Since the consumer neither has the time or knowhow to explore the all movies. A recommendation engine becomes a very important tool to make sure the users of our system get the best experience from the platform.

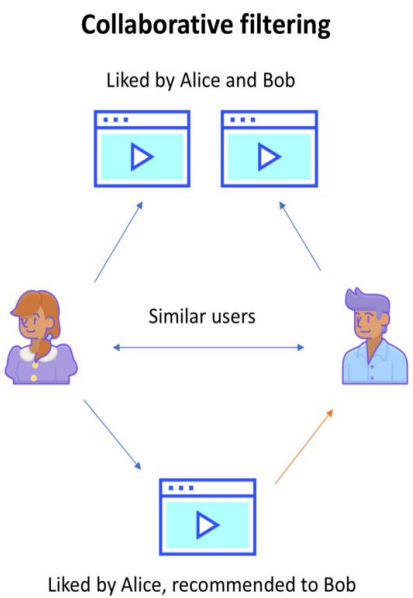
Introduction

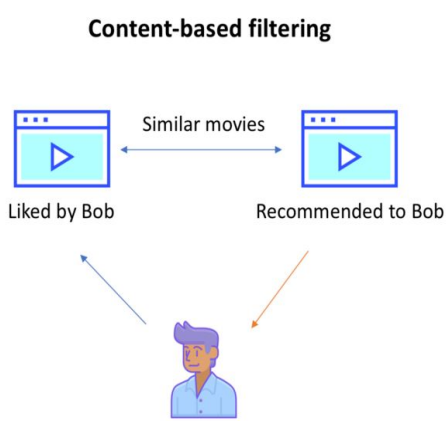
There are various methods for recommendation system -

- Popularity model
- Content-based filtering
- Collaborative filtering
- Hybrid method (Content-based + Collaborative)

Plan

- Provide a collaborative filtering recommendation system with both K-Means and Matrix factorization approaches.
- Build a Content-based filtering system using tf-idf.
- Compare the result of both techniques and show which gives better recommendations.





Dataset - MovieLens 100K

The data is a 5-star rating and free-text tagging activity from MovieLens.It contains 100836 ratings and 3683 tag applications across 9742 movies.

Exploratory Data Analysis:

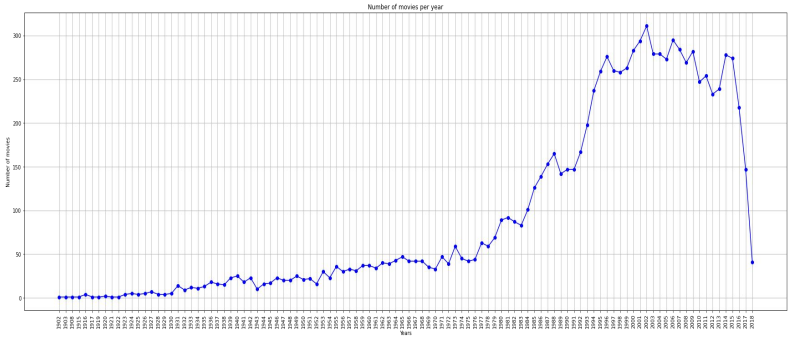
Movies and Years:

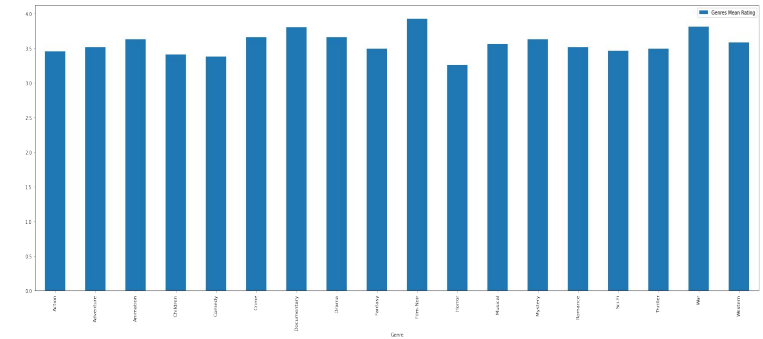
- Contains movies produced from 1902 to 2018.
- The number of movies increased from the year 1985 to 2002 and decreased from 2014 to 2018.

Genres and Ratings:

- The dataset contains contains 18 tags
- The mean rating for the genres is 3.57
- The standard deviation for the rating in the genre is 0.18.

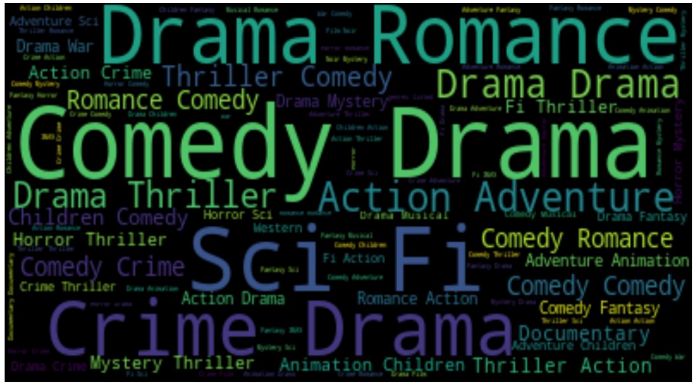
No. of movies per year



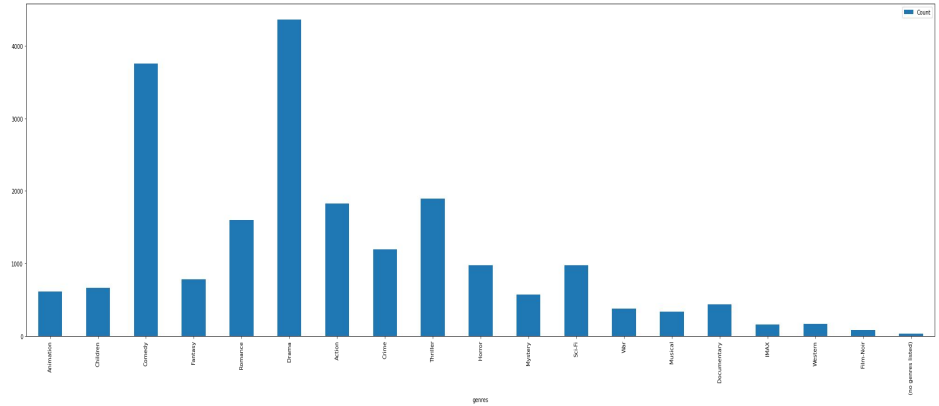


Feature Engineering

- **‘Movie rating’** is very crucial feature. The ratings from different users can be used to identify alike users and recommend movies based on different alike users’ taste.
- Similarly, **‘genre’** can be a helpful feature to build a recommendation engine for movies.
- Dataset has collection of movies classified into 18 different genres. This information can be used to identify list of preferred genres for a user and recommend movies falling into those genres.



Word Cloud for Genre



Count of Movies for each Genre

Collaborative Filtering

Filters information by using the interactions and data collected by the system from other users.

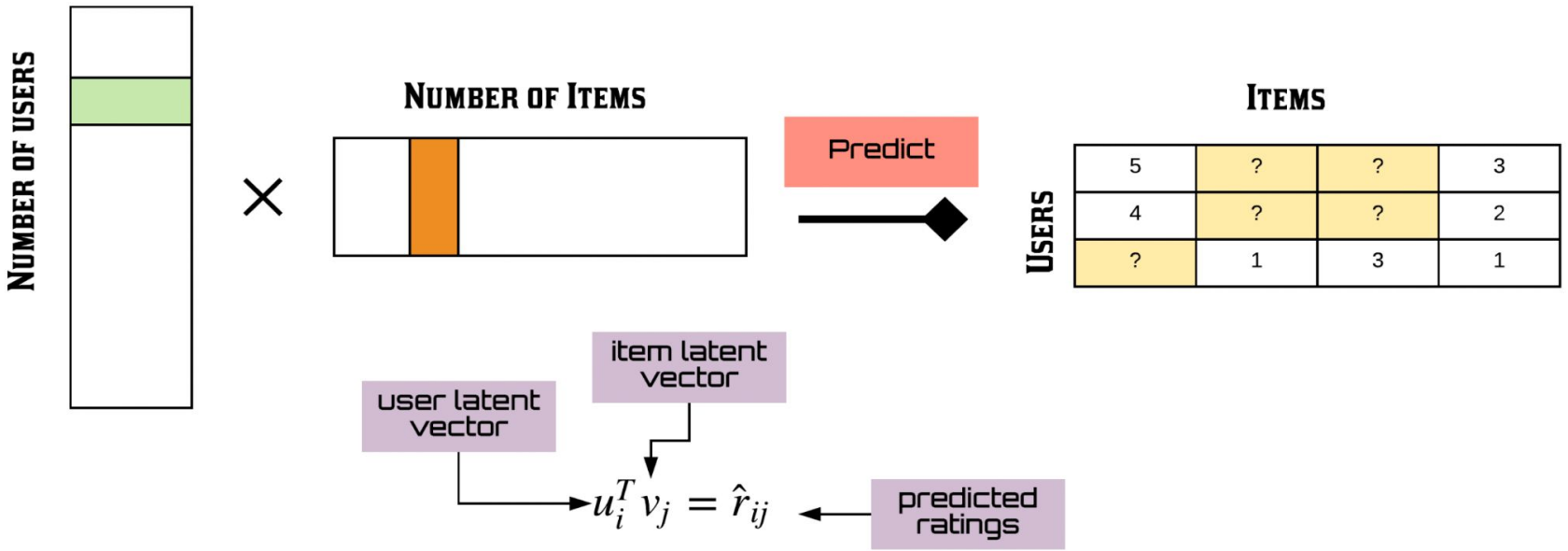
Matrix Factorization generates and finds the best latent features by comparison when multiplying two different kinds of entities.

- **Goal :** Create a matrix of users across movies, where each cell represents the score given by a user for a particular movie.

K-Means Clustering Is an unsupervised learning technique which groups data into clusters based on the similarity between different data-points.

- **Goal :** FInd optimum number of classes and cluster the datasets based on the number of classes

Matrix Factorization



Matrix Factorization

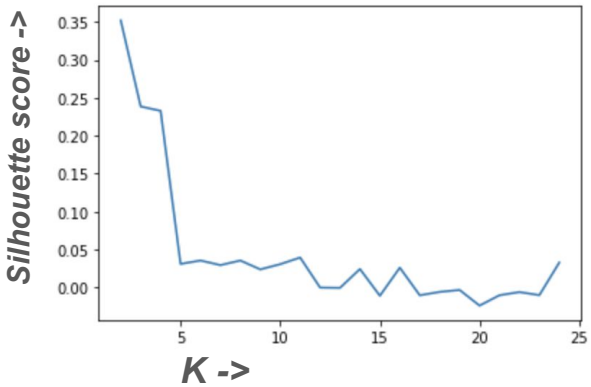
Steps:

1. Split the dataset into training and test samples randomly in a 80/20 ratio.
2. We use the ***SVD*** library in the ***scikit-surprise*** package in python, to decompose user and movies data into latent feature matrices, which predict the top 5 users and the top 5 movies for each user.
3. We use this decomposed data, to predict the ratings for each user using Linear Regression, thereby finding the best latent features.
4. Additionally, we use ***XGBoost*** to improve our prediction accuracy.
5. With this ***XGBoost*** regression data, we try to predict the top 10 movies that a user might like, which is shown in the next slide.

K-means Clustering

Steps:

1. Transform the dataset into a matrix of User x Movies where each cell is a movie rating
   Matrix[i,j] = Rating given by User i to Movie j
2. Since the dataset is very sparse, to fix this, we create a dense region on top of the dataset
   a. Sorted by most rated movies
3. Convert the data into a Compressed Sparse Row matrix
4. Using the Elbow method, find the optimal K
5. Create Clusters



K-means Clustering

Optimal K: To find the optimal number of clusters, we used the Silhouette method:

- **Silhouette Method (**Measures how similar an object is to its own cluster (**cohesion**) compared to other clusters (**separation**))
  - Compute a(i): The average distance of that point with all other points in the same clusters.
  - Compute b(i): The minimum average distance of that point with all the points in the closest cluster to its cluster.

Optimal K found = 6

$$s(i) = \frac{b(i) - a(i)}{\max(b(i), a(i))}$$

Content-Based Filtering

- We used SKlearn's ***TfidfVectorizer*** to vectorize the genres of a movie in combinations of 3.
- Create a grid with all movies mapped with each other
- Using cosine similarity, compare the movies to the input movies and find how similar they are based on their multiple genres
- After which, return the top 10 results as the recommended movies.

Content-Based Filtering

The formula ***TfidfVectorizer*** uses to vectorize the genre data is given below

- Less common genres receive more weight (df_x is the number of occurrences of the genre)
- Combination of 3 genres (or lesser) are considered as a dimension during vectorization.

$$w_{x,y} = tf_{x,y} \times \log\left(\frac{N}{df_x}\right)$$

TF-IDF

Term x within document y

$tf_{x,y}$ = frequency of x in y

df_x = number of documents containing x

N = total number of documents

Result

Collaborative Filtering :

- We display the predictions for User with ID 31 in our dataset using both Matrix Factorization and K-Means clustering.
- We then show the evaluation results for both the methods of Collaborative Filtering.
- The Matrix factorization, we use the user ids, movie and ratings to predict the latent features, using which we predict the top 10 movie recommendations for each user.
- The K-Means process, recommends the top 10 movies in the cluster for each user.

Content Filtering :

- Based on the observation of what the user likes, we find and recommend movies in the same genres of movies he/she like.
- We use the generate the similarity scores for all movies using ***TfidfVectorizer*** and cosine product.
- Then we sum the scores for the movies that the user has rated 5.0 and generate our list of 10 recommendations.

Matrix Factorization Recommendations

Ground truth for user ID: 31

rating	timestamp	title	genres
5.0	850466616	Toy Story (1995)	Adventure|Animation|Children|Comedy|Fantasy
5.0	850466616	Mission: Impossible (1996)	Action|Adventure|Mystery|Thriller
5.0	850466810	Mars Attacks! (1996)	Action|Comedy|Sci-Fi
5.0	850466810	Preacher's Wife, The (1996)	Drama
5.0	850466810	Paradise Lost: The Child Murders at Robin Hood...	Documentary
5.0	850467485	Fantasia (1940)	Animation|Children|Fantasy|Musical
5.0	850467425	Young Frankenstein (1974)	Comedy|Fantasy
5.0	850467408	Princess Bride, The (1987)	Action|Adventure|Comedy|Fantasy|Romance
5.0	850467468	Die Hard (1988)	Action|Crime|Thriller
5.0	850467425	Casablanca (1942)	Drama|Romance

Recommendations for User ID: 31

	movieId	title	genres
405	479	Judgment Night (1993)	Action|Crime|Thriller
254	303	Quick and the Dead, The (1995)	Action|Thriller|Western
68	83	Once Upon a Time... When We Were Colored (1995)	Drama|Romance
492	586	Home Alone (1990)	Children|Comedy
482	569	Little Big League (1994)	Comedy|Drama
275	326	To Live (Huozhe) (1994)	Drama
366	434	Cliffhanger (1993)	Action|Adventure|Thriller
387	458	Geronimo: An American Legend (1993)	Drama|Western
299	352	Crooklyn (1994)	Comedy|Drama
186	229	Death and the Maiden (1994)	Drama|Thriller

K-Means Recommendations

Ground truth for user ID: 31

rating	timestamp	title	genres
5.0	850466616	Toy Story (1995)	Adventure|Animation|Children|Comedy|Fantasy
5.0	850466616	Mission: Impossible (1996)	Action|Adventure|Mystery|Thriller
5.0	850466810	Mars Attacks! (1996)	Action|Comedy|Sci-Fi
5.0	850466810	Preacher's Wife, The (1996)	Drama
5.0	850466810	Paradise Lost: The Child Murders at Robin Hood...	Documentary
5.0	850467485	Fantasia (1940)	Animation|Children|Fantasy|Musical
5.0	850467425	Young Frankenstein (1974)	Comedy|Fantasy
5.0	850467408	Princess Bride, The (1987)	Action|Adventure|Comedy|Fantasy|Romance
5.0	850467468	Die Hard (1988)	Action|Crime|Thriller
5.0	850467425	Casablanca (1942)	Drama|Romance

Recommendations for User ID: 31

movieId	title	genres	score
1197	Princess Bride, The (1987)	Action|Adventure|Comedy|Fantasy|Romance	4.529412
318	Shawshank Redemption, The (1994)	Crime|Drama	4.444444
593	Silence of the Lambs, The (1991)	Crime|Horror|Thriller	4.296296
1136	Monty Python and the Holy Grail (1975)	Adventure|Comedy|Fantasy	4.285714
1196	Star Wars: Episode V - The Empire Strikes Back...	Action|Adventure|Sci-Fi	4.272727
608	Fargo (1996)	Comedy|Crime|Drama|Thriller	4.239130
1193	One Flew Over the Cuckoo's Nest (1975)	Drama	4.178571
1270	Back to the Future (1985)	Adventure|Comedy|Sci-Fi	4.173913
1356	Star Trek: First Contact (1996)	Action|Adventure|Sci-Fi|Thriller	4.104167
1968	Breakfast Club, The (1985)	Comedy|Drama	4.076923

Content-based Recommendations

Ground truth for user ID: 31

rating	timestamp	title	genres
5.0	850466616	Toy Story (1995)	Adventure|Animation|Children|Comedy|Fantasy
5.0	850466616	Mission: Impossible (1996)	Action|Adventure|Mystery|Thriller
5.0	850466810	Mars Attacks! (1996)	Action|Comedy|Sci-Fi
5.0	850466810	Preacher's Wife, The (1996)	Drama
5.0	850466810	Paradise Lost: The Child Murders at Robin Hood...	Documentary
5.0	850467485	Fantasia (1940)	Animation|Children|Fantasy|Musical
5.0	850467425	Young Frankenstein (1974)	Comedy|Fantasy
5.0	850467408	Princess Bride, The (1987)	Action|Adventure|Comedy|Fantasy|Romance
5.0	850467468	Die Hard (1988)	Action|Crime|Thriller
5.0	850467425	Casablanca (1942)	Drama|Romance

Recommendations for User ID: 31

	movieId	title	genres
4	5	Father of the Bride Part II (1995)	Comedy
17	18	Four Rooms (1995)	Comedy
18	19	Ace Ventura: When Nature Calls (1995)	Comedy
58	65	Bio-Dome (1996)	Comedy
61	69	Friday (1995)	Comedy
79	88	Black Sheep (1996)	Comedy
90	102	Mr. Wrong (1996)	Comedy
92	104	Happy Gilmore (1996)	Comedy
104	119	Steal Big, Steal Little (1995)	Comedy
108	125	Flirting With Disaster (1996)	Comedy

Results: Matrix Factorization

Training Accuracy: (Collaborative Filtering)

Method	**Regression - RMSE**	**Regression - MAPE**	**XGBoost - RMSE**	**XGBoost - MAPE**
Matrix Factorization	0.7698768620174847	25.0212126398069	0.7612402213604075	24.50911485864543

Testing Accuracy:

Method	**Regression - RMSE**	**Regression - MAPE**	**XGBoost - RMSE**	**XGBoost - MAPE**
Matrix Factorization	0.706339965920172	21.861796658407794	0.7714926455791599	23.938527922400954

Results: K-Means

Method	**RMSE**	**MAPE**
K-Means	0.8134992095683429	25.154572013297656

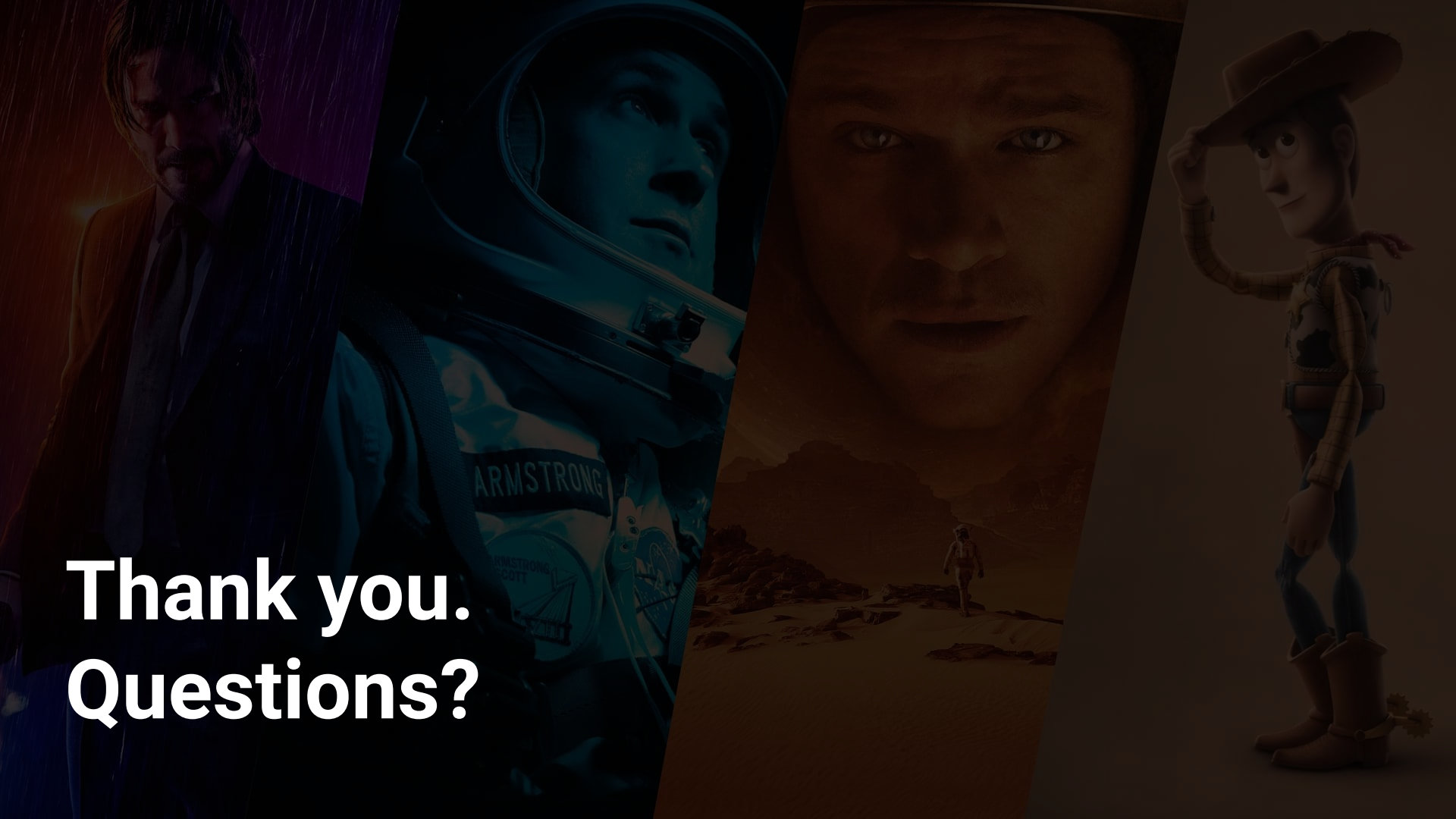
Analysis

- During training, XGBOOST had a lower RMSE score than Regression when it comes to Matrix factorization. However, after testing, we observed that the Regression had performed better than XGBOOST.

- In general Matrix Factorization(MF) performed better than K-means clustering. It can be observed that MF testing values for RMSE was lower(better) than K-means.

- Although both the methods have a good error score, MAPE less than 25 and RMSE below 2, MF had a significantly lower MAPE score than K-means clustering.

- Content Based filtering also was able to successfully predict movies in the same genre and similar to the preference of users



Thank you.
Questions?