

Movie Recommendation System

Team Members: Nancy Jain(MT18041), Abhay Agrawal(MT18123), Sakshi Arora(MT18133)

- **AIM:** In this project, we aim to give personalized recommendations to users by preparing various models based on movies' ratings. We focused on three main algorithms: Baseline(Generalised Recommender), Content-Based Recommender and Collaborative Filtering.
- **COMPLEXITY OF DATA ACQUISITION:** The dataset is obtained from kaggle - The Movie Dataset, curated from Full MovieLens Dataset. The dataset comprises of 100,000 ratings and 1,300 tag applications applied to 9,000 movies by 700 users. Ratings are on a scale of 1-5 and have been obtained from the official GroupLens website. Our project considers a subset of the original data comprising of details of 4320 movies and ratings given by 6000 users. Since the dataset is quite large, we have combined useful data(genres, cast, crew, etc) from different CSV files that are useful in movie recommendation.
- ❖ **Data Cleaning:** As the dataset is quite large, many of the entries are empty, hence the very first step is to handle the missing values.

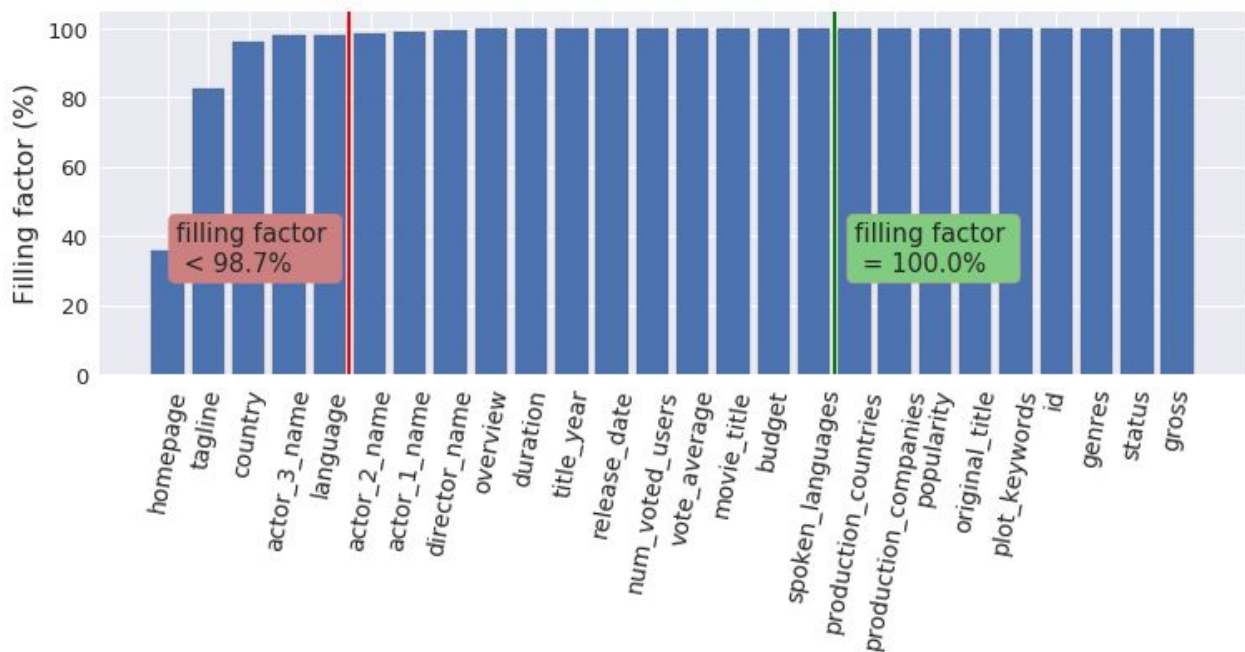


Fig 1 : Analysis of missing data in the dataset

❖ Dataset Visualization and Analysis :



Fig 2 : Most popular genres in the dataset

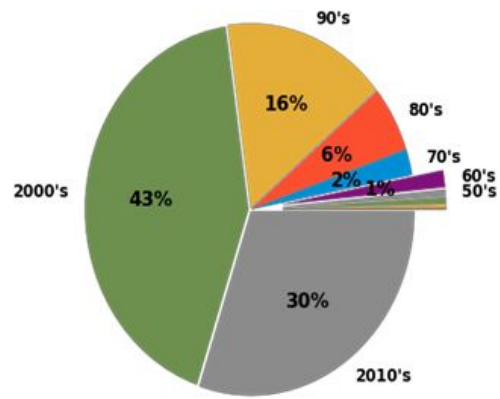


Fig 3 : Number of films per year

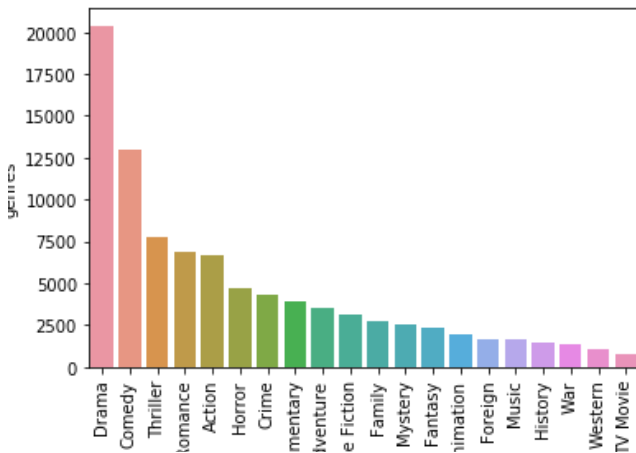


Fig 4 : Analysing genre data

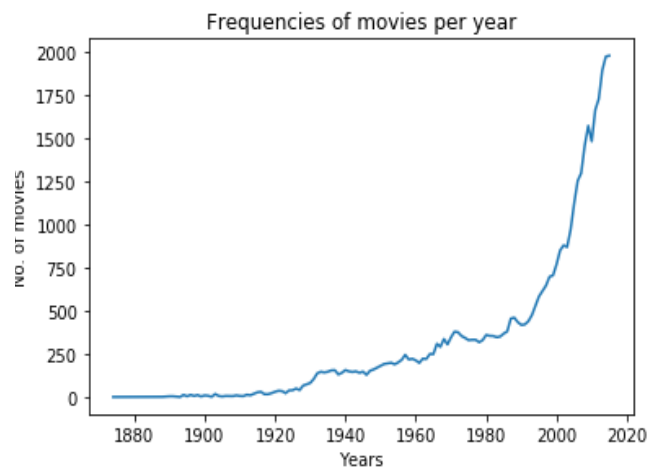


Fig 5 : Exponential increase in number of movies over years

- **HYPOTHESES:** Most of the state-of-the-art recommender systems often fail to grasp the nature of users choice. Therefore, we attempt to develop models that cover both aspects: content-based as well as user-specific recommendations through different types of rating mechanisms.

● PREDICTIVE MODELS USED :

1. Generalized Recommendation Model: It gives the same recommendation to each user based on movie popularity or genre. The idea is that the movies which are popular and critically acclaimed have high chances of being liked by an audience.

2. Content-Based Recommendation Model: The suggestions are given on the basis of a particular movie. If a user is currently watching movie X, then X's metadata like genre, actors, description, keywords etc are used to further suggest movies. This recommender system works on the assumption that if a user liked movie X, then he would also like movies similar to X.

3. Collaborative Filtering based Recommendation System: This technique is used to personalise recommendations to a user based on his interests. CF techniques basically find a correlation between user's interests and movie's features and on its basis recommends a personalised movie.

There are 2 types of Collaborative Filtering Techniques:

User-Based: Here we assume that similar people have similar liking. If user A and user B have watched a lot of similar movies and have rated them almost equally, then it can be inferred that their interests match. Thus we can recommend movies that A has liked to B.

Item-Based: Here we take into consideration the content of movies as well and not only user similarity. Here we find those movies which the user would enjoy based on his previously rated movies.

● VISUALIZATION AND STATISTICS TO SUPPORT HYPOTHESES :

➤ Results of Generalised Recommendation Model and Content-Based Recommendation Model :

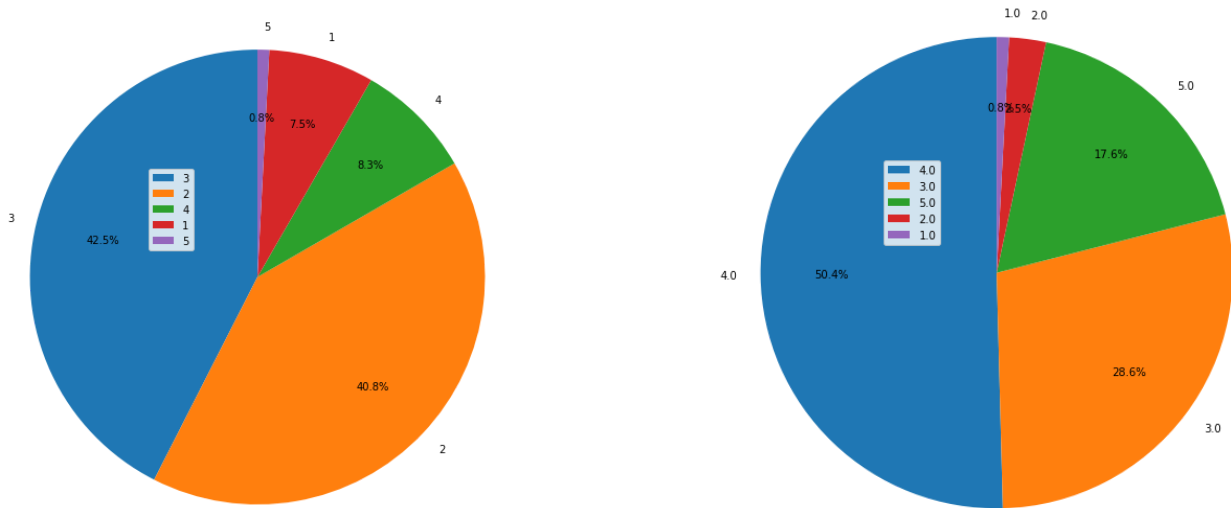


Fig 6 : Customer Feedback on Generalised Recommendation and Content Based Recommendation system

➤ Results of Collaborative Filtering Based Recommendation Model :

Features	Test MAE	Test RMSE	Train MAE	Train RMSE
Genre, Crew	0.6829436039	0.8827476799	0.7599656357	0.9556038375
Crew, Keywords	0.8418156809	1.036022444	0.9125429553	1.097067318
Genre ,Crew ,Production Countries	0.7138927098	0.9162599656	0.776975945	0.9657538053
Genre, Production Countries, Keywords	0.6777166437	0.8672158173	0.7374570447	0.9352398299
Genre, Crew, Production Countries, Keywords	0.6905089409	0.8905587631	0.7451890034	0.9341828485

Table 1 : RMSE and MAE values on 80 train and 20% test data



Fig 8 : Comparison of scores on train and test data

- **CONCLUSION** : The observed results show that among all the feature combinations selected, *Genres*, *Production countries*, & *keywords* play an important role in rating a movie. The performance of the applied item-based collaborative filtering is found to be best among all three methods as can be observed by RMSE scores on test data.
- **REPRODUCIBILITY** : In contrast to the state-of-the-art systems where the suggestion is delivered from a content perspective, this new model delivers a more user-centred recommender system. User Based Collaborative Filtering is a very common and highly used approach for movielens dataset because of the fact that:
 - It is easy to implement.
 - It is context independent.
 - It gives better results than content based recommendation system.

But this approach is not the best. Recommending movies without taking into consideration the content of the movie would never give best results.

Thus we tried to implement an **item based recommendation system** with our movies dataset.

- **REFERENCES :**

- [1] Xiaoyuan Su and Taghi M. Khoshgoftaar, "A Survey of Collaborative Filtering Techniques," *Advances in Artificial Intelligence*, vol. 2009, Article ID 421425, 19 pages, 2009.
- [2] Koren, Yehuda. "Factor in the neighbours: Scalable and accurate collaborative filtering." *ACM Transactions on Knowledge Discovery from Data (TKDD)* 4.1 (2010): 1.
- [3] <https://realpython.com/build-recommendation-engine-collaborative-filtering/>