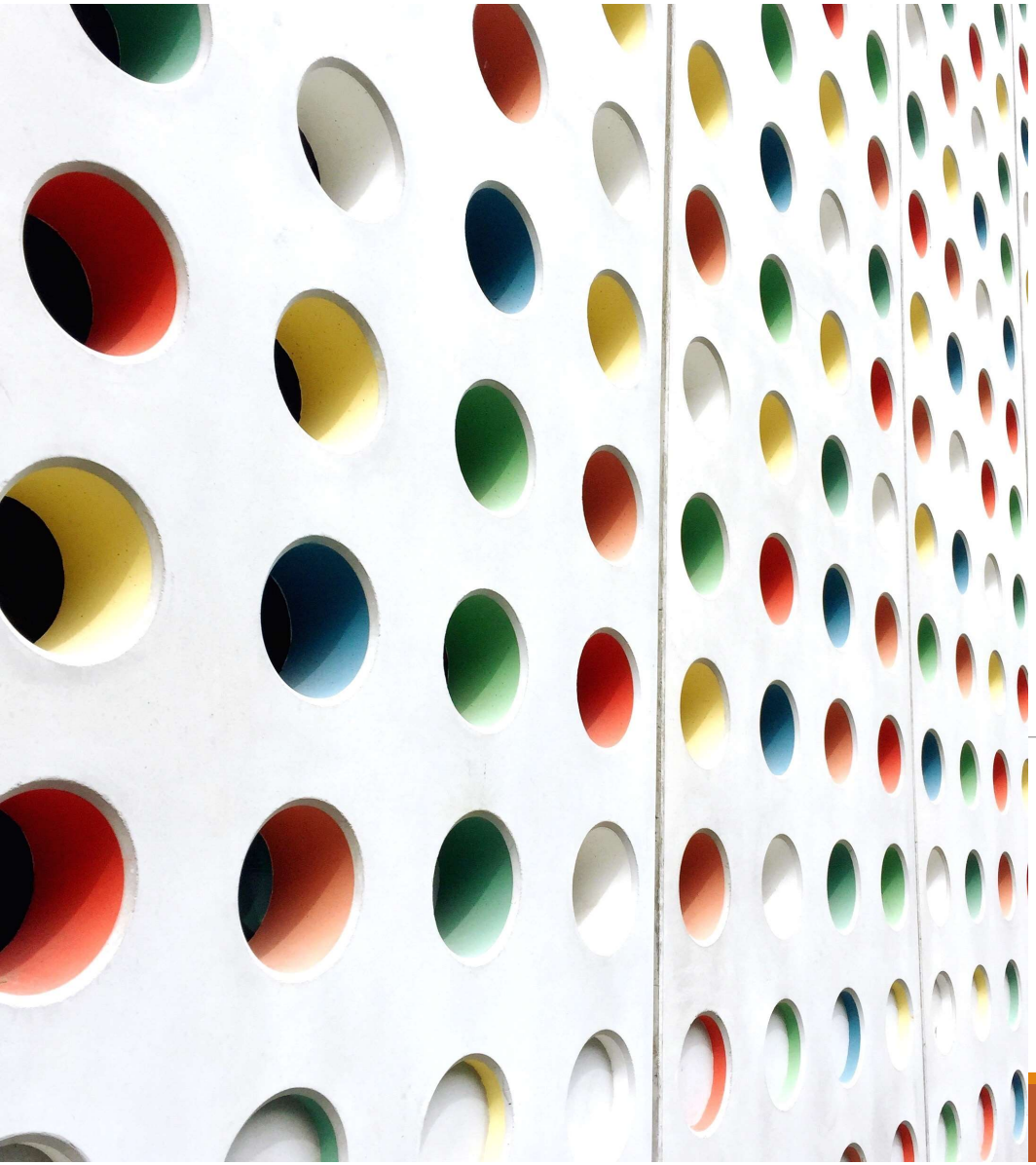


MOVIE RECOMMENDATION SYSTEM

DATA SCIENCE CAPSTONE PROJECT

Springboard

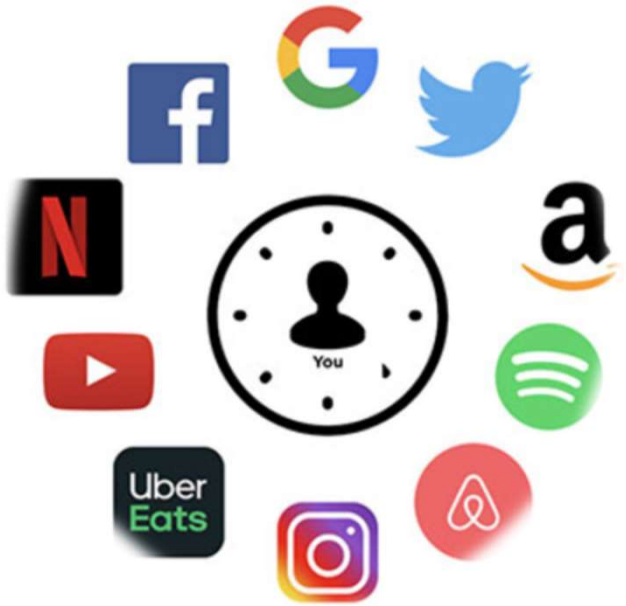
BY,

LUBNA RAHMAN

Recommendation System: An Overview

- Recommendation System is an information filtering system.
- Predicts the rating or the preference a user might give to an item.
- Is ubiquitous because of abundant data.
- Growing in popularity in the world of predictive modeling and Machine Learning.
- Netflix, YouTube, Facebook, and Amazon are some popular recommender systems in use.

Recommendation system: Influencing everyday lives



Project Goals

- Understanding, analyzing and correlating the trend in movie ratings per genre.
- Building a recommender engine to provide recommendations to different users.

Project Overview

Building a movie recommendation system to provide recommendations to number of users with numerous movies.

Recommendation is done based on :

- Similarity between users (Collaborative Filtering), or
- Considering particular user's activity (Content Based Filtering) which they want to engage with.

Business Objective

- The recommender engine is useful to the businesses that earns revenue via recommendations.
- Providing good recommendations will help users save time searching for their favorite movie.
- This will help the customer to continue with the service.
- Clients of this project could be Amazon, Netflix, Hulu, YouTube.

Workflow

The recommendation system project workflow includes:

- Data collection
- Data Wrangling
- Exploratory Data Analysis(EDA)
- Pre-processing and Modeling

Data Collection

- Data Source: https://grouplens.org/datasets/movielens/25m/
- Obtained from MovieLens, an online recommendation service.
- The dataset has two dataframes, the ‘ratings’ dataset and the ‘movies’ dataset.
- The ‘ratings’ dataset consists of around 25 million user ratings
- The movies dataset provides information about the movie name, genre and the year of release.

Data Wrangling

- The dataset was clean and had no missing values.
- The 'ratings'dataset has 25,000,095 rows and 4 columns, and
- The 'movies'dataset has 62,423 rows and 3 columns.
- The 'genres' column was separated from the movies dataset and merged with the ratings dataset for better data exploration.
- The 'year' part was removed from the 'title' column and,
- 'year' and 'Decade' were made as separate columns.

Ratings and Movies Dataset

Ratings Dataset: Rows and Columns

```
1 r.shape
```

```
(25000095, 4)
```

	userId	movieId	rating	timestamp
0	1	296	5.0	1147880044
1	1	306	3.5	1147868817
2	1	307	5.0	1147868828
3	1	665	5.0	1147878820
4	1	899	3.5	1147868510

Movies Dataset: Rows and Columns

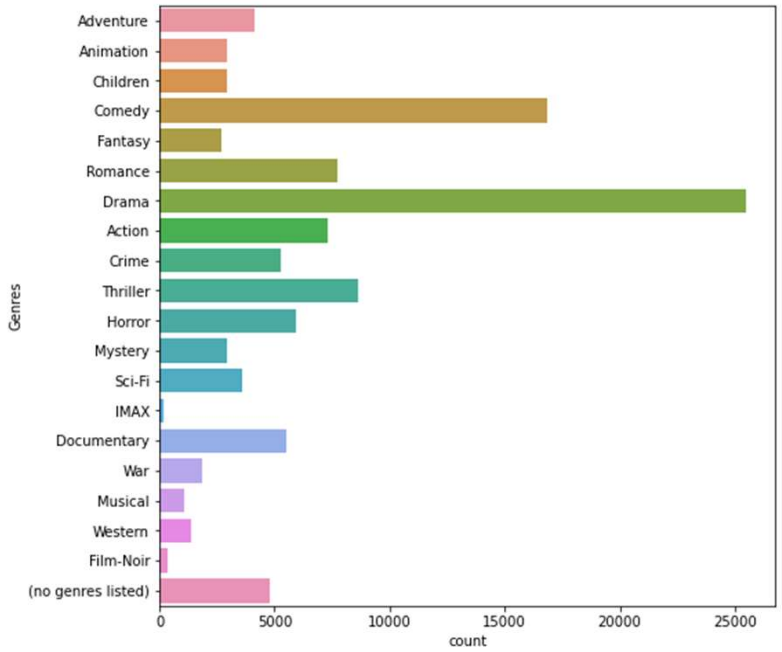
```
1 movies.shape
```

```
(62423, 3)
```

movieId	title	genres
1	Toy Story (1995)	Adventure|Animation|Children|Comedy|Fantasy
2	Jumanji (1995)	Adventure|Children|Fantasy
3	Grumpier Old Men (1995)	Comedy|Romance
4	Waiting to Exhale (1995)	Comedy|Drama|Romance
5	Father of the Bride Part II (1995)	Comedy

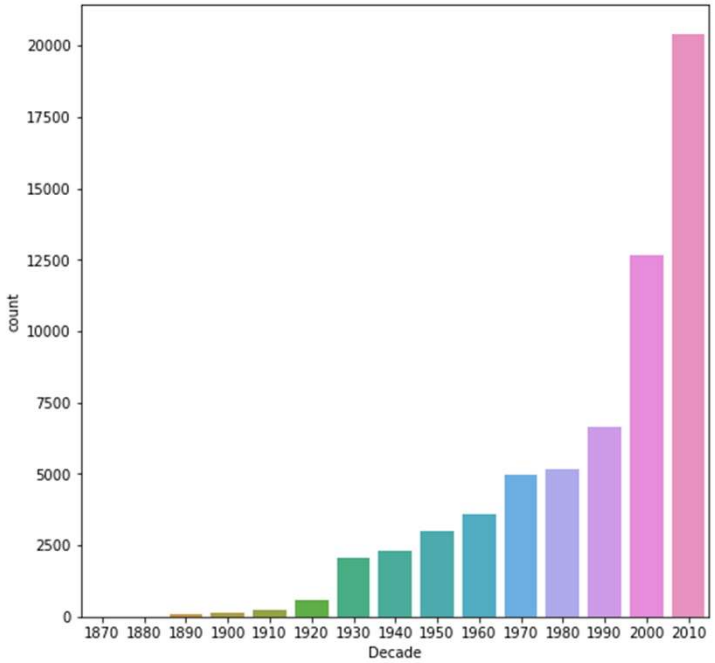
Exploratory Data Analysis

1. Total number of movies per each Genre: ***Drama and Comedy*** have the most production followed by ***Thriller, Romance and Action***.



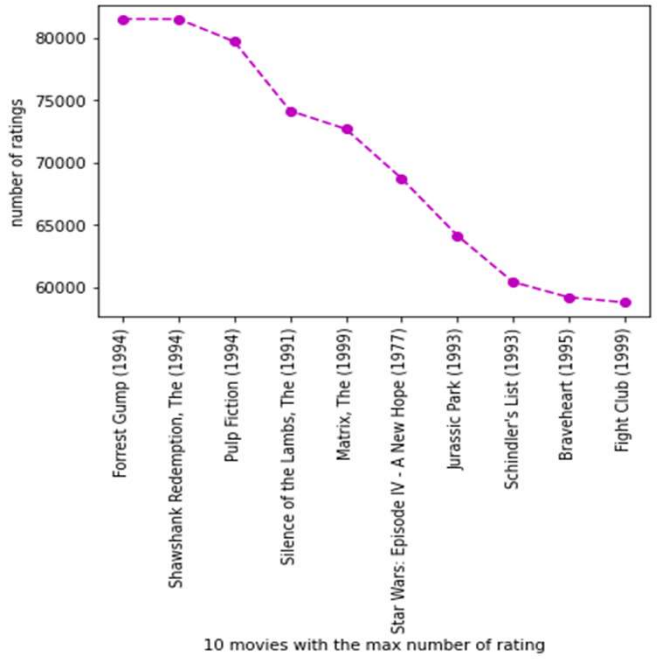
Exploratory Data Analysis

Total number of movies in each decade: The plot shows an increasing trend with most of the production being in the **2010** decade.



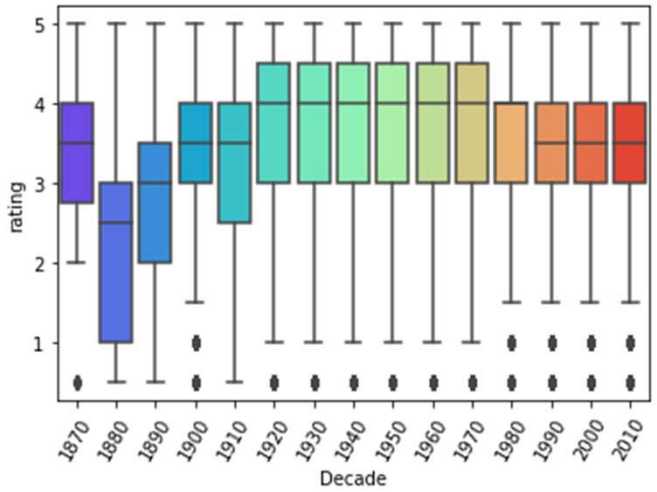
Top 10 Movies

The top 10 movies with highest ratings are: ***Forrest Gump, Shawshank Redemption and Pulp Fiction .***



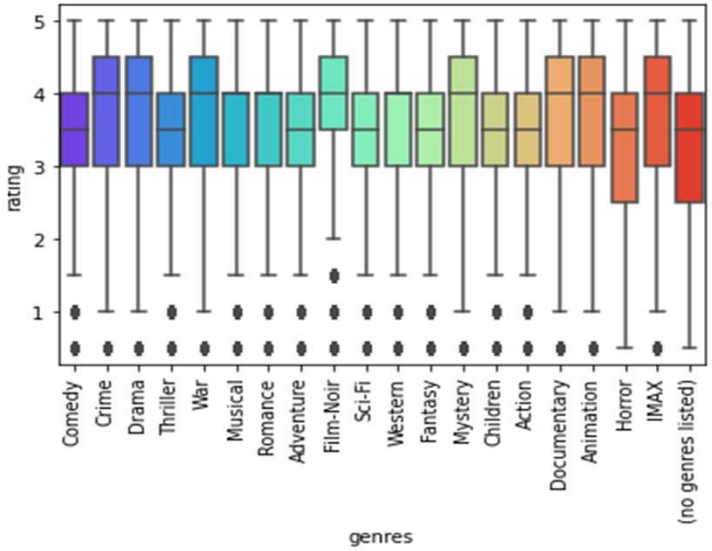
Rating Vs Decade

The boxplot below shows that the movies of the decade spanning from 1920 to 1970 have the highest ratings.



Rating Vs Genres

The boxplot shows that ***Drama***, ***Crime***, ***War***, ***Mystery*** are some of the most highly rated genres while ***Horror*** is the least favorite.



Word Cloud

The word cloud for movie ‘Titles’ : The most common word used in movie titles is ***"Love".***

Modeling

1. Simple Recommender: It generalizes the recommendations to users based on movie popularity, ratings and genre. Hence, popular and highly acclaimed movies will have higher probability of being liked by the audience.

2. IMDb weighted rating formula: It publishes weighted vote averages rather than just raw data averages. In other words, although we accept and consider all votes received by users, not all votes have the same impact (or 'weight') on the final rating.

3. Content based filtering: It uses item features to recommend other items similar to what the user likes, based on their previous actions or explicit feedback.

4. Collaborative Filtering: It uses algorithms to filter data from user reviews to make personalized recommendations for users with similar preferences. For this project ***Matrix Factorization*** based collaborative filtering algorithms have been used.

Matrix Factorization

The different matrix factorization based algorithms used in this project are:

- Singular Value Decomposition(SVD)
- SVD++
- LightFM
- Alternative Least Square (ALS) collaborative Filtering

Singular Value Decomposition(SVD)

- SVD is a matrix factorization technique that is used to reduce the number of features of a dataset by reducing the space dimensions from N to K, where K< N.
- For the recommendation systems, we will focus on the matrix factorization keeping the same dimensionality. It is done on the user-item ratings matrix.
- ***Surprise library*** was used which consists of several powerful algorithms like SVD, NMF(Non-negative Matrix Factorization). KNN(K- Nearest Neighbor) and Co-Clustering, to minimize the RMSE(Root Mean Square Error) and give accurate recommendations.

SVD Observations

SVD

```
Evaluating RMSE, MAE of algorithm SVD on 5 split(s).

                  Fold 1  Fold 2  Fold 3  Fold 4  Fold 5  Mean    Std
RMSE (testset)    0.8704  0.8726  0.8656  0.8610  0.8678  0.8675  0.0040
MAE (testset)     0.6671  0.6696  0.6644  0.6606  0.6642  0.6652  0.0030


SVD
Total Elapsed time with model is: 7.859
Prediction of a particular user: 4.32803974336067
Actual rating value is: 5.0

Prediction of a particular user: 4.096602008912272
Actual rating value is: 3.5

Prediction of a particular user: 4.167597853757669
Actual rating value is: 4.0
```

SVD++

- As the results from the SVD model did not improve even after hyperparameter tuning, we tried the model ***SVD++.***
- SVD++ algorithm is an extension of SVD that takes into account the implicit ratings. In other words, it as an optimized SVD algorithm to enhance the accuracy of prediction by generating implicit feedback.
- However, SVD++ also did not give us any good results. The RMSE for this algorithm was still found to be around 0.65.

LightFM Model

- LightFM is a Python implementation of a number of popular recommendation algorithms for both implicit and explicit feedback.
- It is a hybrid matrix factorisation model representing users and items as linear combinations of their content features' latent factors.
- It also makes it possible to incorporate both item and user metadata into the traditional matrix factorization algorithms.
- It represents each user and item as the sum of the latent representations of their features, thus allowing recommendations to generalise to new items (via item features) and to new users (via user features).
- The LightFM model gave us the precision at K score of around 0.053 and an AUC score of 0.91, which is quite good. To further improve the model, we tried Hyperparameter Tuning.
- Hyperparameter Tuning, improves the model performance by roughly around 3%. Hence, this model is considered a good recommender system for this dataset.

Alternating Least Square(ALS)

- ALS is also a matrix factorization algorithm, which factorizes a given matrix R into two factors U and V such that R≈UTV.
- The unknown row dimension is given as a parameter to the algorithm and is called latent factors.
- AIS is an iterative optimization process where for every iteration, we move closer to an factorized representation of our original data.
- We will use this model to fit our data and find the similarities.
- It was observed that the ALS model gives good scores and hence, good recommendations for this dataset.

ALS: Observations

1. Movies similar to Shawshank Redemption

	MovieId	Score	Title
0	318	1.000000	Shawshank Redemption, The (1994)
1	527	0.998930	Schindler's List (1993)
2	593	0.997405	Silence of the Lambs, The (1991)
3	50	0.996919	Usual Suspects, The (1995)
4	356	0.996794	Forrest Gump (1994)
5	296	0.996069	Pulp Fiction (1994)
6	110	0.993827	Braveheart (1995)
7	47	0.989299	Seven (a.k.a. Se7en) (1995)
8	480	0.986144	Jurassic Park (1993)
9	293	0.985520	Léon: The Professional (a.k.a. The Professiona...

2. Movies similar to Forrest Gump

	MovieId	Score	Title
0	356	1.000000	Forrest Gump (1994)
1	593	0.996942	Silence of the Lambs, The (1991)
2	527	0.996888	Schindler's List (1993)
3	318	0.996794	Shawshank Redemption, The (1994)
4	110	0.994411	Braveheart (1995)
5	296	0.992466	Pulp Fiction (1994)
6	50	0.992437	Usual Suspects, The (1995)
7	480	0.991438	Jurassic Park (1993)
8	47	0.983593	Seven (a.k.a. Se7en) (1995)
9	1	0.980937	Toy Story (1995)

THANK YOU!