Movie Recommendation Capstone Project

Rohit Goyal

4/4/2021

1. Introduction

1.1 Objective

The objective of this project is to develop a movie recommendation model that will predict a rating for a given user and movie combination. Movie Recommendation models are used in online video streaming platforms such as Netflix and Amazon Prime Video to provide relevant suggestions to customers.

1.2 Dataset Used

The "movielens" dataset (https://grouplens.org/datasets/movielens/10m/) will be used to build the recommendation model using machine learning. The dataset contains 10 million movie rating records and was released in 2009. Along with **rating** given by a user for a **movie**, the data set also contains the **timestamp** of when a rating was given as well as the **genres** tagged to a particular movie.

1.3 Summary of Steps

- Prepare the the movielens data set for analysis
- Split the dataset into 2 parts - training set (to build the recommendation model) and validation set (to check the performnce of model)
- Explore the dataset to identify trends that will form the basis for building the model
- Incrementally build the recommendation model
- Assess performance of the final model on the validation dataset

2. Detailed Analysis of Steps

2.1 Data Preparation

- The first step is to download the movielens data set zip file from the website. Once downloaded, the ratings and movie info data are read from 2 separate files.
- The ratings and movie info data is then combined together to get the movielens data set. This dataset is then split into 2 parts.

1. 90% as training set (to build the model)
2. Remaining 10% as validation set (to test our final model).

- It is further ensured that users and movies in validation set are present in the training set.
- Next, the year in which a rating was given is derived from the timestamp attribute.

2.2 Data Exploration and Visualization

- The data is now ready for initial exploration. Here is a snippet of the prepared data.

userId	movieId	rating	timestamp	title	genres	rate_year
1	122	5	838985046	Boomerang (1992)	Comedy|Romance	1996
1	185	5	838983525	Net, The (1995)	Action|Crime|Thriller	1996
1	292	5	838983421	Outbreak (1995)	Action|Drama|Sci-Fi|Thriller	1996
1	316	5	838983392	Stargate (1994)	Action|Adventure|Sci-Fi	1996
1	329	5	838983392	Star Trek: Generations (1994)	Action|Adventure|Drama|Sci-Fi	1996
1	355	5	838984474	Flintstones, The (1994)	Children|Comedy|Fantasy	1996

- The dimesion of dataset (rows, columns) and the total distinct users and movies are as follows:

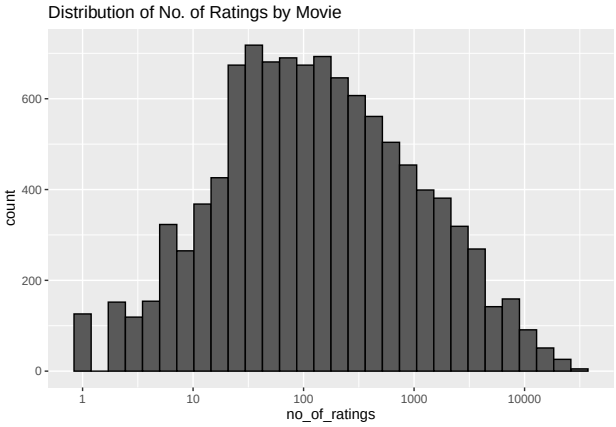
Dimensions: rows, columns
9000055
7

distinct_users	distinct_movies
69878	10677

- The top 10 movies with highest number of ratings are shown below and as expected these are popular and well recognized movies.

movieId	no_of_ratings	title
296	31362	Pulp Fiction (1994)
356	31079	Forrest Gump (1994)
593	30382	Silence of the Lambs, The (1991)
480	29360	Jurassic Park (1993)
318	28015	Shawshank Redemption, The (1994)
110	26212	Braveheart (1995)
457	25998	Fugitive, The (1993)
589	25984	Terminator 2: Judgment Day (1991)
260	25672	Star Wars: Episode IV - A New Hope (a.k.a. Star Wars) (1977)
150	24284	Apollo 13 (1995)

- A frequency distribution of no of ratings by movie is plotted (log scale) and we see that majority of the movies have very low number of ratings. Only the popular movies tend to attract ratings in large numbers.



- The most frequently given ratings are checked and it is found that majority 4 and 3 are the most top most given ratings. Whole number ratings are given more than fractional ratings.

rating	rating_count
4.0	2588430
3.0	2121240
5.0	1390114
3.5	791624
2.0	711422
4.5	526736
1.0	345679
2.5	333010
1.5	106426
0.5	85374

- The top 10 genres with most number of ratings show that Drama and Comedy movies are the most rated ones.

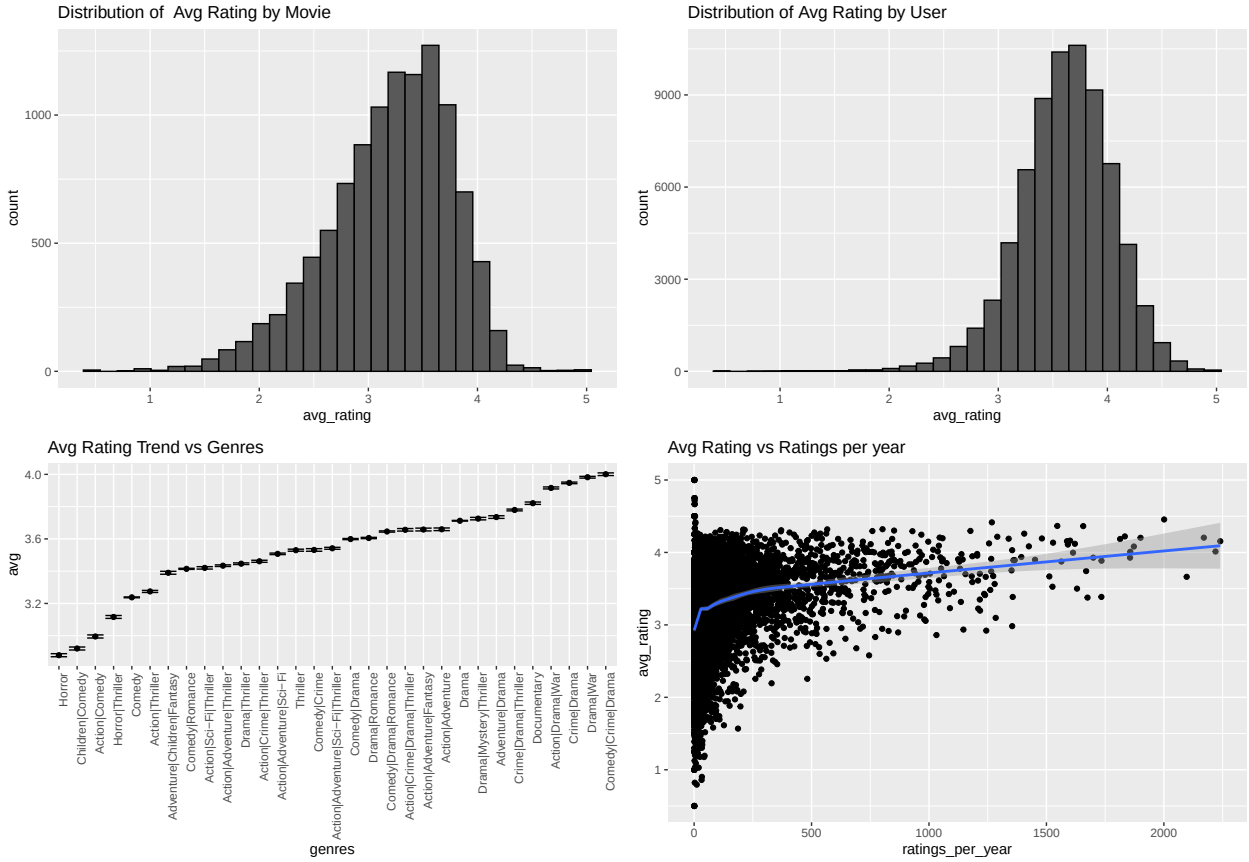
genres	genres_count
Drama	733296
Comedy	700889
Comedy|Romance	365468
Comedy|Drama	323637
Comedy|Drama|Romance	261425
Drama|Romance	259355
Action|Adventure|Sci-Fi	219938
Action|Adventure|Thriller	149091
Drama|Thriller	145373
Crime|Drama	137387

- Avg Rating variability by different attributes are expolored and following observations are made - These form the basis of subsequent model development by provding potential factors that are important to prediction of ratings for a given user and movie.

1. **By Movie** - There is sufficient variability by movie as expected. The no of movies falls progressively

as one moves towards the tails (i.e. <3 and >4) of the distribution.

2. **By Users** - Again as common sense would dictate, users have unique preferences and rating styles and therefore good variability observed across users - from very picky users to very generous users.
3. **By Genres** - Some genres tend to attract better ratings than others as observed. E.g. Drama movies have higher average ratings while horror movies get lower average ratings.
4. **By ratings per year** - While there is no clear trend for movies which are not rated frequently, the consensus starts to emerge for movies which are rated more frequently.



2.3 Incremental Model Development

- The training set is first further split into training and test data sets for model development.
- RMSE function will be used to check model performance
- **Model 1** - Use simple avg rating of all movies across all users. As expected this simplistic model does not give a great performance.

Model	RMSE
Simple average	1.059904

- **Model 2** - In this increment, the movie effect is added to the model. As expected, an improvement is seen in performance.

Model	RMSE
Simple average	1.0599043
Movie Effect Model	0.9437429

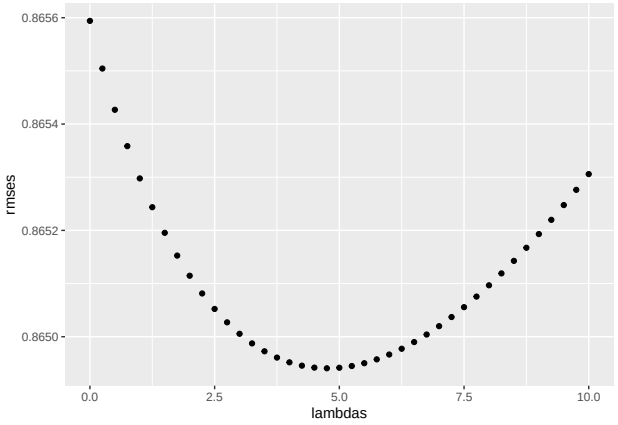
- **Model 3** - In this increment, the user effect is added to the model. Again, this improves the model further as seen by reduction in RMSE.

Model	RMSE
Simple average	1.0599043
Movie Effect Model	0.9437429
Movie + User Effects Model	0.8659320

- **Model 4** - In this increment, the genres effect is added to the model. This improves the model even further as seen by reduction in RMSE. Though the improvement is not as much as seen in earlier model increments.

Model	RMSE
Simple average	1.0599043
Movie Effect Model	0.9437429
Movie + User Effects Model	0.8659320
Movie + User+ Genres Effects Model	0.8655941

- **Model 5** - As a final enhancement to the model, regularization technique is applied and the tuning paramter lambda is calculated. A plot of RMSE for different values of lambda is shown below and a minimum is achieved for a value of 4.75. The RMSE for our final model is 0.8649406, a good benefit achieved with regularization as shown in the table of RMSE for our incremental models.



Regularziation lambda
4.75

Model	RMSE
Simple average	1.0599043
Movie Effect Model	0.9437429
Movie + User Effects Model	0.8659320
Movie + User+ Genres Effects Model	0.8655941
Movie + User+ Genres Effects + Regularization Model	0.8649406

- Before checking the performance on Validation dataset, the final model is now retrained on the entire Training set using the chosen lambda. This because the Training set had been split into further subsets thus far during model development.

3. Model Performance with Validation Data Set

- The final model performance is now tested on the Validation data set, i.e the final hold-out set.
- As shown below, the RMSE score achieved is 0.8644514 which is pretty close to what we saw for our final model with training data (which was 0.8649406).

Final Model RMSE with Validation Set
0.8644514

4. Conclusion

- The stated objective of building a movie recommendation model is achieved as presented in this report.
- The model is built using the movielens dataset by analysing trends in the training data which give insight on what factors affect the rating. The identified factors of movie, user and genre are incrementally incorporated into the final model and regularization technique is used to further enhance the final model.
- The final model is tested on validation dataset and the model performance is consistent with what is observed on training data. Infact the performance is found to be slightly better on validation data.

4.1 Limitations and future work

- The model could be enhanced to recommend better ratings for cases where a user or a movie is not present in the training data set. For this project we specifically ensured that validation set did not have such cases.
- The final proposed model did not include rating frequency even though it was identified as an informative factor during data exploration. This could be analysed further.
- Other matrix based machine learning techniques such as Principal component analysis can also be explored to see if they can help build better models.