Movie Lens Report

Sulaiman Saleh AlAwad

15 Jun 2019

Create edx set, validation set, and submission file

MovieLens 10000054 dataset from
(http://files.grouplens.org/datasets/movielens/ml-10m.zip)

Executive Summary

The purpose for this project is creating a recommender system using MovieLens dataset.

The version of movielens dataset used for this final assignment contains approximately 10 Million of movies ratings, divided in 9 Million for training and one Million for validation. It is a small subset of a much larger (and famous) dataset with several millions of ratings. Into the training dataset there are approximately 70.000 users and 11.000 different movies divided in 20 genres such as Action, Adventure, Horror, Drama, Thriller and more.

After a initial data exploration, the recommender systems built on this dataset are evaluated and chosen based on the RMSE - Root Mean Squared Error that should be at least lower than 0.87750

For accomplishing this goal, the Regularized Movie+User+Genre Model is capable to reach a RMSE of 0.8628, that is really good.

The RMSE function that will be used in this project is:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n} e_t^2}$$

Exploratory Data Analysis

Initial data Exploration

The 10 Million dataset is divided into two datasets: **edx** for training purpose and **validation** for the validation phase.

The edx dataset contains approximately 9 Million of rows with 70.000 different users and 11.000 movies with rating score between 0.5 and 5. There is no missing values (0 or NA).

edx dataset

	User Id	Movie Id	rating	timestamp	title	genres
1	1	122	5	838985046	Boomerang (1992)	Comedy|Romance
2	1	185	5	838983525	Net, The (1995)	Action|Crime|Thriller
3	1	231	5	838983392	Dumb & Dumber (1994)	Comedy
4	1	292	5	838983421	Outbreak (1995)	Action|Drama|Sci-Fi|Thriller
5	1	316	5	838983392	Stargate (1994)	Action|Adventure|Sci-Fi
6	1	329	5	838983392	Star Trek: Generations (1994)	Action|Adventure|Drama|Sci-Fi

The features/variables/columns in both datasets are six:

1. userId as integer: that contains the unique identification number for each user.
2. movieId as umeric: that contains the unique identification number for each movie.
3. rating as numeric: that contains the rating of one movie by one user. Ratings are made on a 5-Star scale with half-star increments.
4. timestamp as integer: that contains the timestamp for one specific rating provided by one user.
5. title as character: that contains the title of each movie including the year of the release.
6. genres as character: that contains a list of pipe-separated of genre of each movie.

Dataset Pre-Processing and Feature Engineering

After an initial data exploration, we notice that the genres are pipe-separated values. It's necessary to extract them for more consistent, robust and precise estimate. We also observe that the title contains the year where the movie war released and this it could be necessary to predic the movie rating. Finally, we can extract the year and the month for each rating.

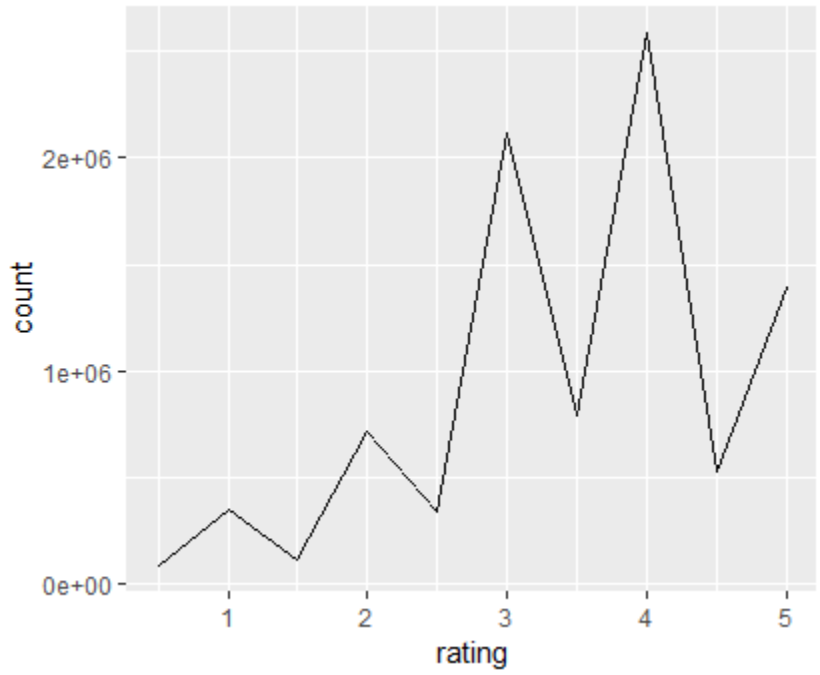
The pre-processing phase is composed by these steps:

1. Convert timestamp to a human readable date format;
2. Extract the month and the year from the date;
3. Extract the release year for each movie from the title;
4. Separate each genre from the pipe-separated value. It increases the size of both datasets.

After preprocessing the data, edx dataset looks like this:

Overview of Rating Distribution

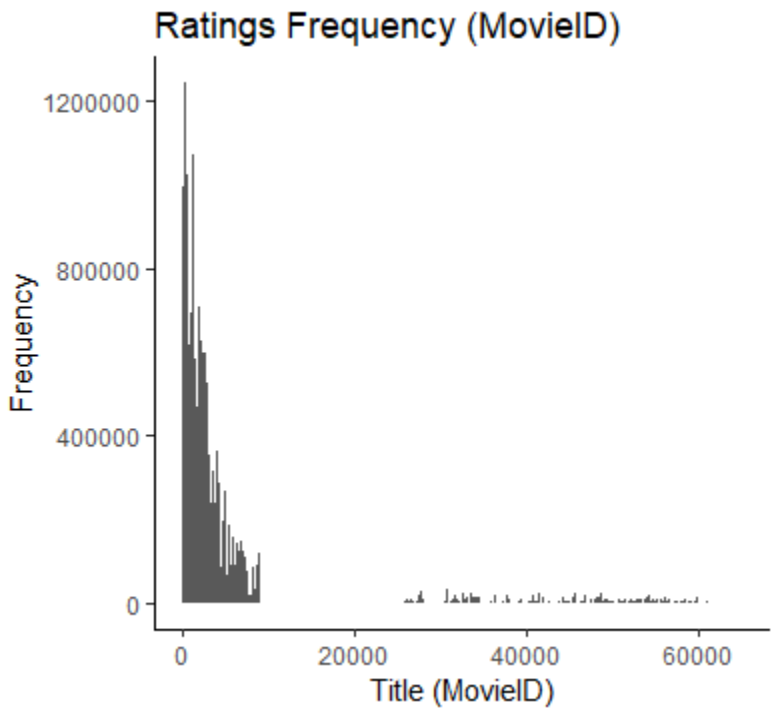
According to the histogram below, it shows that there are a small amount of negative votes (below 3). Maybe, the user tends to give a vote if he liked the movie. Half-Star votes are less common than "Full-Star" votes.



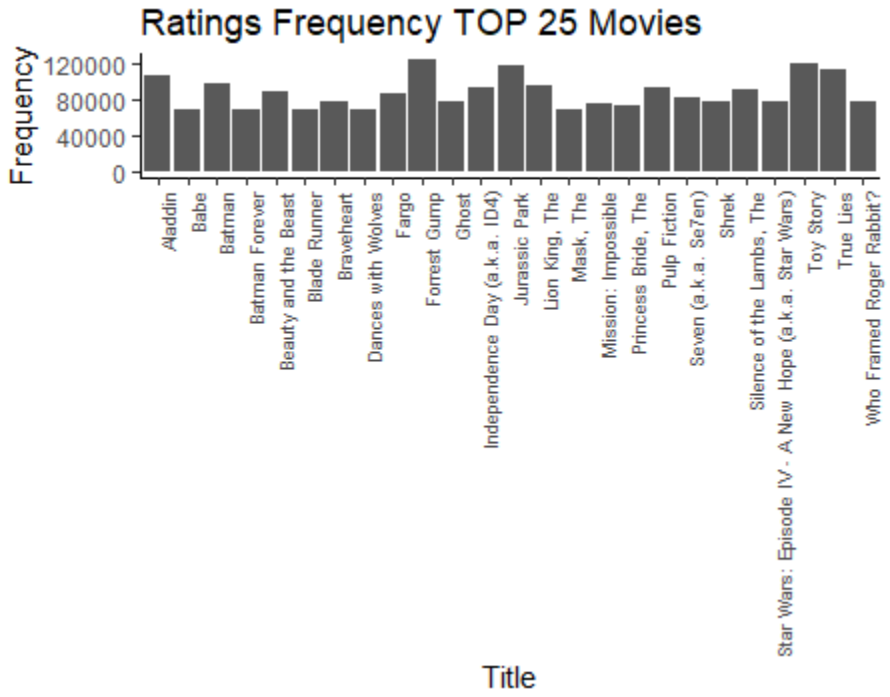
In table:

No	ratings	Number of given ratings
1	zeros	0
2	ones	345935
3	tows	710998
4	threes	2121638
5	fours	2588021
6	fives	1390541

Numbers of Ratings per Movie

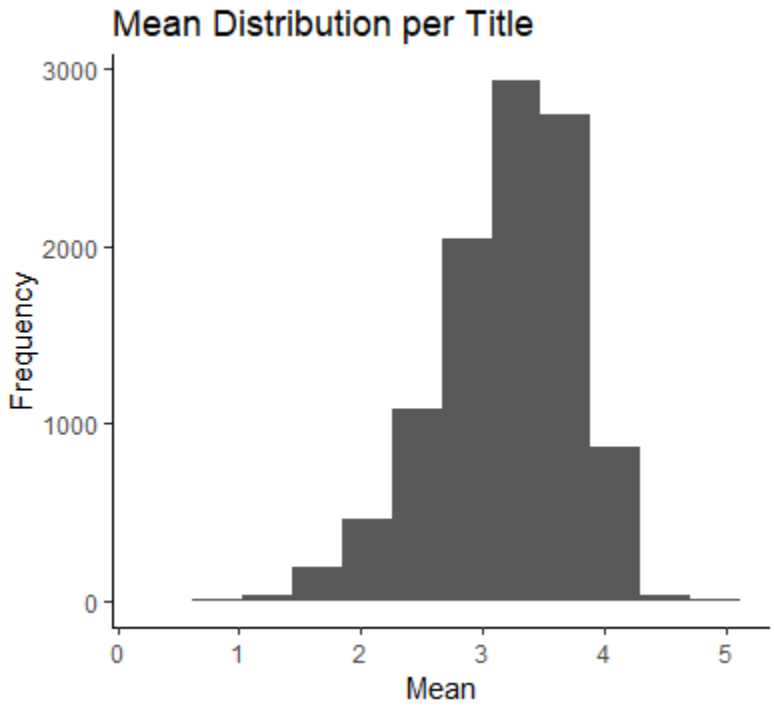


Top Rated Movies

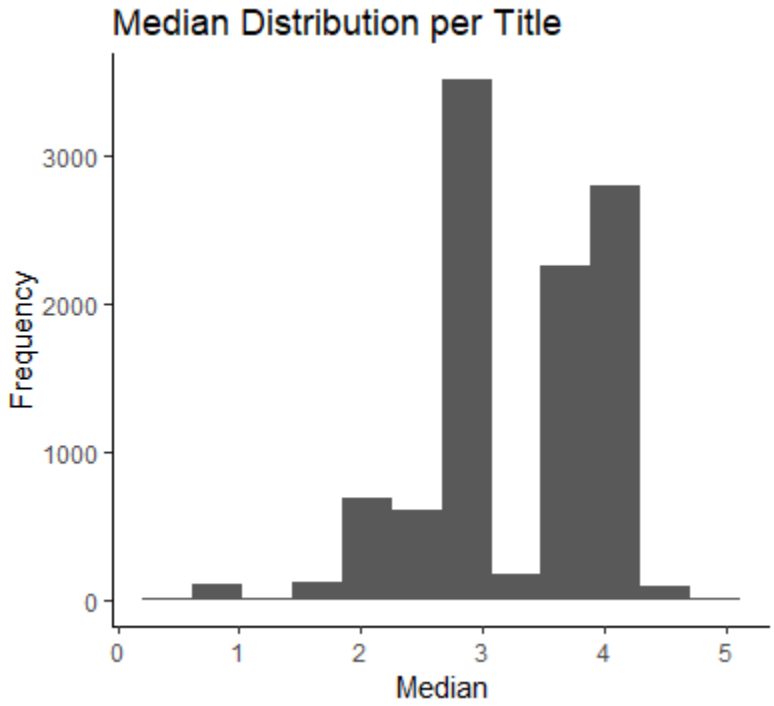


No.	movieId	title	count
1	296	Pulp Fiction (1994)	31336
2	356	Forrest Gump (1994)	31076
3	593	Silence of the Lambs, The (1991)	30280
4	480	Jurassic Park (1993)	29291
5	318	Shawshank Redemption, The (1994)	27988
6	110	Braveheart (1995)	26258
7	589	Terminator 2: Judgment Day (1991)	26115
8	457	Fugitive, The (1993)	26050
9	260	Star Wars: Episode IV - A New Hope (a.k.a. Star Wars) (1977)	25809
10	592	Batman (1989)	24343

Mean Distribution per Title (Movie ID)



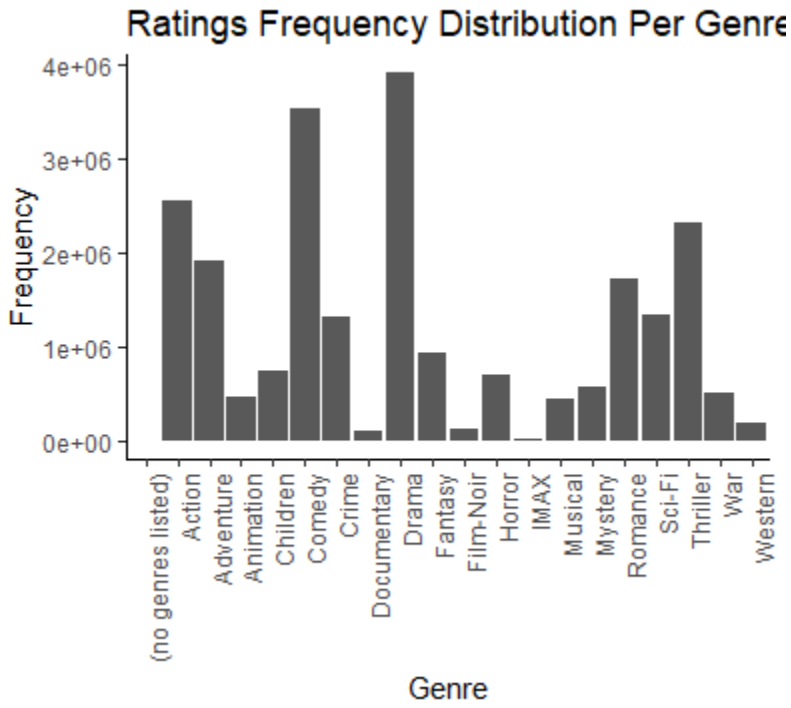
Median Distribution per Title (Movie ID)



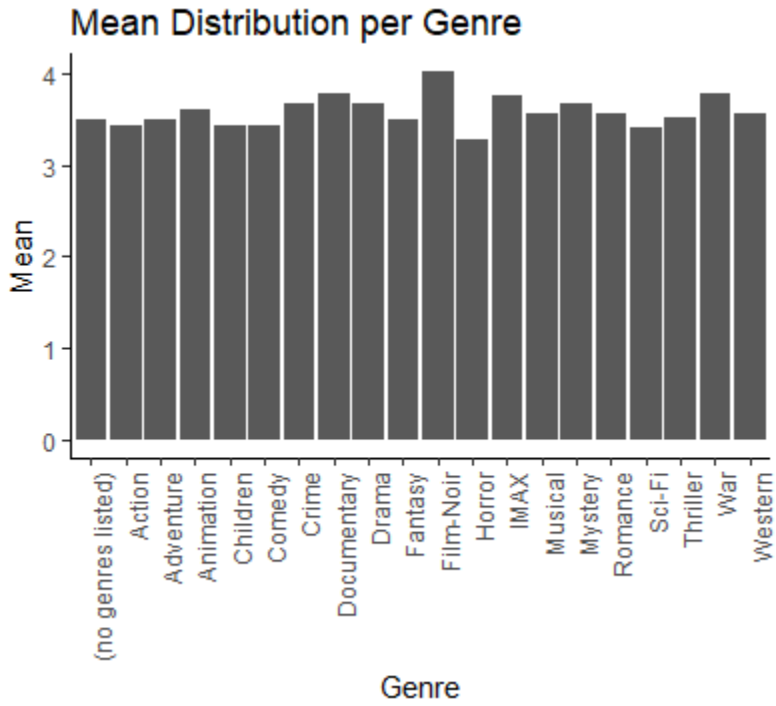
Genre Analysis

Rating Distribution per Genre

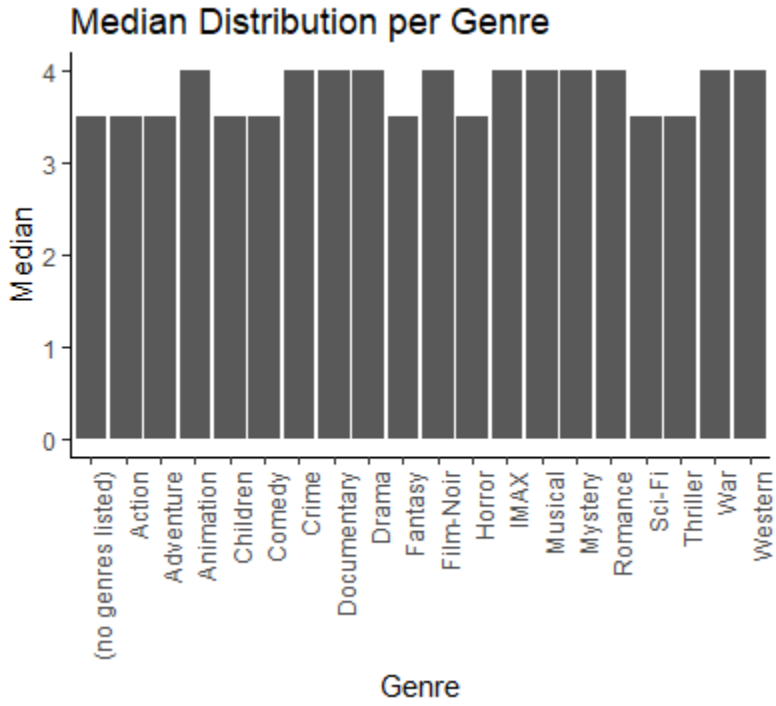
Overview of Rating distribution over Genre



Mean Distribution per Genre



Median Distribution per Genre



Analysis - Model Building and Evaluation

Naive Baseline Model

The simplest model that someone can build, is a Naive Model that predict ALWAYS the mean. In this case, the mean is approximately 3.5.

The mean is: 3.527004

The RMSE on the validation dataset is 1.05. It is very far for the target RMSE (below 0.87) and that indicates poor performance for the model.

Movie-Based Model, a Content-based Approach

The RMSE on the validation dataset is 0.94. It better than the Naive Mean-Baseline Model, but it is also very far from the target RMSE (below 0.87) and that indicates poor performance for the model.

Movie + User Model, a User-based approach

The RMSE on the validation dataset is 0.8635 and this is very good. The Movie+User Based Model reaches the desidered performance but applying the regularization techniques, can improve the performance just a little.

Movie + User + Genre Model, the Genre Popularity

The RMSE on the validation dataset is 0.8634 and this is very good. The Movie+User+Genre Based Model reaches the desidered performance but adding the genre predictor, doesn't improve significantly the model's performance.

Applying the regularization techniques, can improve the performance just a little.

Regularization

The RMSE on the validation dataset is 0.8635 and this is very good. The Movie+User Based Model reaches the desidered performance but applying the regularization techniques, can improve the performance just a little.

Regularized Movie+User Model

The RMSE on the validation dataset is 0.8629 . The Regularized Movie+User Based Model improves just a little the result of the Non-Regularized Model.

Regularized Movie+User+Genre Model

The RMSE on the validation dataset is 0.8628 and this is the best result of the builted models. The Regularized Movie+User+Genre Based Model improves just a little the result of the Non-Regularized Model. As the Non-Regularized Model, the genre predictor doesn't improve significantly the model's performance.

Results

This is the summary results for all the model builted, trained on edx dataset and validated on the validation dataset.

```
results
                          model      RMSE
1    Naive Mean-Baseline Model 1.0524433
2            Movie-Based Model 0.9411063
3       Movie+User Based Model 0.8635899
4 Movie+User+Genre Based Model 0.8634946
```

Conclusion

After training different models, it's very clear that movieId and userId contribute more than the genre predictor. Without regularization, the model can archieves and overtakes the desidered peformance, but the best is the enemy of the good and applying regularization and adding the genre predictor, it make possible to reach a RSME of 0.8628 that is the best result for the trained models.