A Recommendation System for Movies

Joe King

22 June, 2020

Introduction

The purpose of this project is to provide a machine learning model for a movie recommendation system. The data on which this model is produced is the Movielens 10M dataset available from GroupLens research lab

The Movielens 10M dataset comprises around 10 million ratings categorised by user, movie and one or more genres. Approximately 10% of the data is reserved as a validation set, with the remainder being used to train and test the modelling of the recommendation system. Despite being a partitition of the full datset, the validation set will be treated as "unseen" as far as developing the recommendation model is concerned. It will only be used at the end of the process to validate the accuracy of the model. The allocation of the Movielens 10M dataset to training/testing and to validation is shown in Table 1.

Table 1: Summary of Movielens 10M Data

Dataset	No of Rows	No of Movies	No of Users
Full Movielens 10M	10000054	69878	10677
Training/Testing (90% of Movielens 10M)	9000055	69878	10677
Validation (10% of Movielens 10M)	999999	68534	9809

The metric for measuring the accuracy of the model produced is the root mean square error (**RMSE**) represented by the equation

$$\textbf{RMSE} = \sqrt{(\tfrac{1}{\textbf{N}}) \sum_{u,i} (\hat{y}_{u,i} - y_{u,i})^2}$$

where $y_{u,i}$ is the rating for movie i by user u, $\hat{y}_{u,i}$ is the corresponding prediction produced by the model, and **N** is is the number of user/movie combinations in the dataset.

The report begins with some analysis of the dataset in order to look for patterns and correlations within the data. The analysis (on the whole of the Training/Testing dataset (**edx**) informs the choices of methods used for modelling the predictions. The investigation looks at movie and user biases (or effects). These are split into two types, namely per-movie and per-user average biases, and then time-dependent movie and user biases.

Next, the details of how each bias is modelled is set out and the recommendation system model is built. The **edx** dataset is first split 90/10 into training and testing datasets **edx_train** and **edx_test** as set out in Table 2.

Table 2: Summary of **edx** Training and Test Datasets

Dataset	No of Rows	No of Movies	No of Users
edx: Training Set	8100067	69878	10677
edx: Test Set	899988	68081	9719

A linear model is used for developing the recommendation system. Each bias modelled is added to the average rating over all movies and all users. Our estimate for the average rating (denoted $\hat{\mu}$) is the average over all the ratings in the **edx** dataset. The predicted rating for a given row in a dataset will then be $\hat{\mu}$ plus the sum of the biases for the parameters in that row of the dataset.

These predictions are derived from the data in the **edx_train** dataset and tested on the **edx_test** dataset.

Finally the recommendations system model is used to derive predicted ratings on the validation dataset and report the results of the **RMSE** calculation.

Data Analysis

Average Per-Movie Ratings

Intuitively, some movies are better than others, so it is to be expected that there is a spread of ratings between movies. By the same token, "blockbuster" movies are likely to be rated more often than say "art house" movies.

Tables 3 and 4 below show the best and worst ten movies by average rating. What is important to notice is these are generally obscure movies with just a handful of ratings, so it may be necessary to prevent these seldom-rated movies from skewing the recommendations model, suggesting regularisation of the data will be necessary.

Table 3: Best 10 Movies By Average Rating

Title	Average Rating	No of Ratings
Hellhounds on My Trail	5.00	1
Satan's Tango (Sátántangó)	5.00	2
Shadows of Forgotten Ancestors	5.00	1
Fighting Elegy (Kenka erejii)	5.00	1
Sun Alley (Sonnenallee)	5.00	1
Blue Light, The (Das Blaue Licht)	5.00	1
Who's Singin' Over There? (a.k.a. Who Sings Over There) (Ko to tamo peva)	4.75	4
Human Condition II, The (Ningen no joken II)	4.75	4
Human Condition III, The (Ningen no joken III)	4.75	4
Constantine's Sword	4.75	2

Table 4: Worst 10 Movies By Average Rating

Title	Average Rating	No of Ratings
Relative Strangers	1.0000000	1
From Justin to Kelly	0.9020101	199
Disaster Movie	0.8593750	32
Hip Hop Witch, Da	0.8214286	14
SuperBabies: Baby Geniuses 2	0.7946429	56
Besotted	0.5000000	2
Hi-Line, The	0.5000000	1
Accused (Anklaget)	0.5000000	1
Confessions of a Superhero	0.5000000	1
War of the Worlds 2: The Next Wave	0.5000000	2

If we now look at the top five movies by their number of ratings (good, bad or indifferent), then we see very much the "usual suspects", as Table 5 shows.

Table 5: Best 5 Movies By No of Ratings

Title	Average Rating	No of Ratings
Pulp Fiction	4.154789	31362
Forrest Gump	4.012822	31079
Silence of the Lambs, The	4.204101	30382
Jurassic Park	3.663522	29360
Shawshank Redemption, The	4.455131	28015

It is interesting to see a scatter plot of the average movie ratings by the number of times they are rated. Figure 1 shows this. The best and worst ten movies by rating are highlighted in red and our the top five movies by number of ratings are highlighted in red.

The pattern clearly shows how as the number of ratings increase, the average rating converges towards the centre. Our best and worst movies by rating are outliers in this plot, the top 5 by ratings count are in the same range as other movies. Another observation is that there is a weak positive correlation between the average rating and the number of ratings (ρ = 0.2114161).

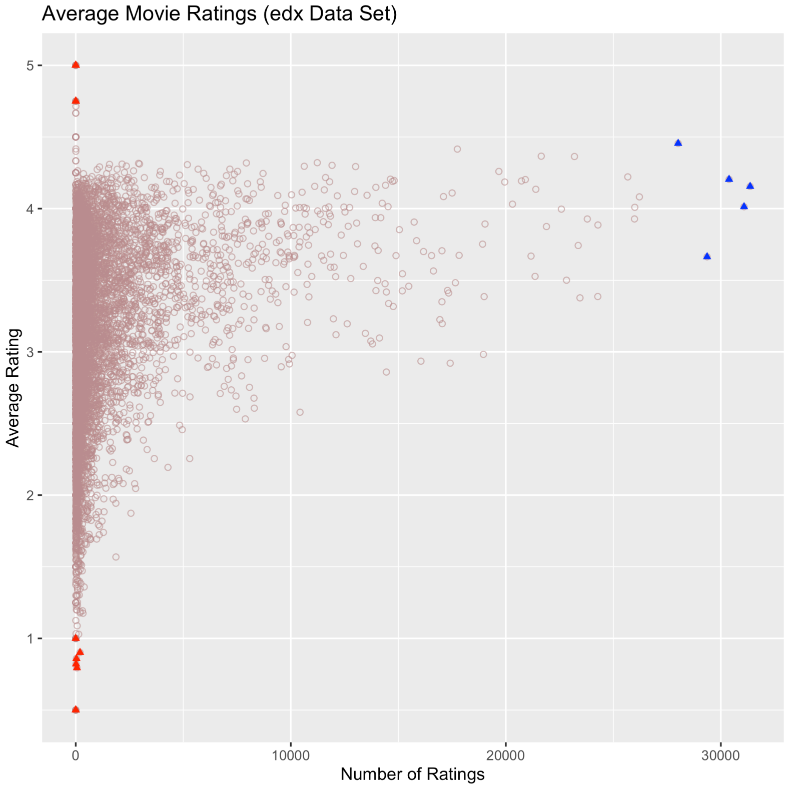

Figure 1: Average Movie Ratings

Average Per-User Ratings

As is the case for movies, it is to be expected that there is a spread of average ratings between users. Similarly, users who are avid movie watchers are likely to rate movies more often than other users. If user tastes are thrown into the mix, then intuitively the variability between users is going to be greater than for movies.

Tables 6 and 7 below show the ten users who give the highest and lowest movie ratings respectively. Table 8 shows the five most prolific raters of movies. As for movies, it's important to notice that generally those high and low ratings are from infrequent raters.

Again, it's important to note that we may need to regularise the data to ensure that the infrequent raters do not skew the model.

Table 6: Users with the 10 Highest Average Ratings

User ID	Average Rating	No of Ratings
1	5	19
7984	5	17
11884	5	18
13027	5	29
13513	5	17
13524	5	20
15575	5	29
18965	5	49
22045	5	18
26308	5	17

Table 7: Users with the 10 Lowest Average Ratings

User ID	Average Rating	No of Ratings
28416	1.0384615	26
3457	1.0000000	19
24176	1.0000000	131
24490	1.0000000	17
6322	0.7058824	17
13496	0.5000000	17
48146	0.5000000	25
49862	0.5000000	17
62815	0.5000000	20
63381	0.5000000	18

Table 8: 5 Most Frequent Rating Users

User ID	Average Rating	No of Ratings
59269	3.264586	6616
67385	3.197720	6360
14463	2.403615	4648
68259	3.576933	4036
27468	3.826871	4023

A scatter plot of the average user ratings by the number of times they rate is shown in Figure 2. The highest and lowest users by their average rating are highlighted in red and the top five most prolific users are highlighted in red.

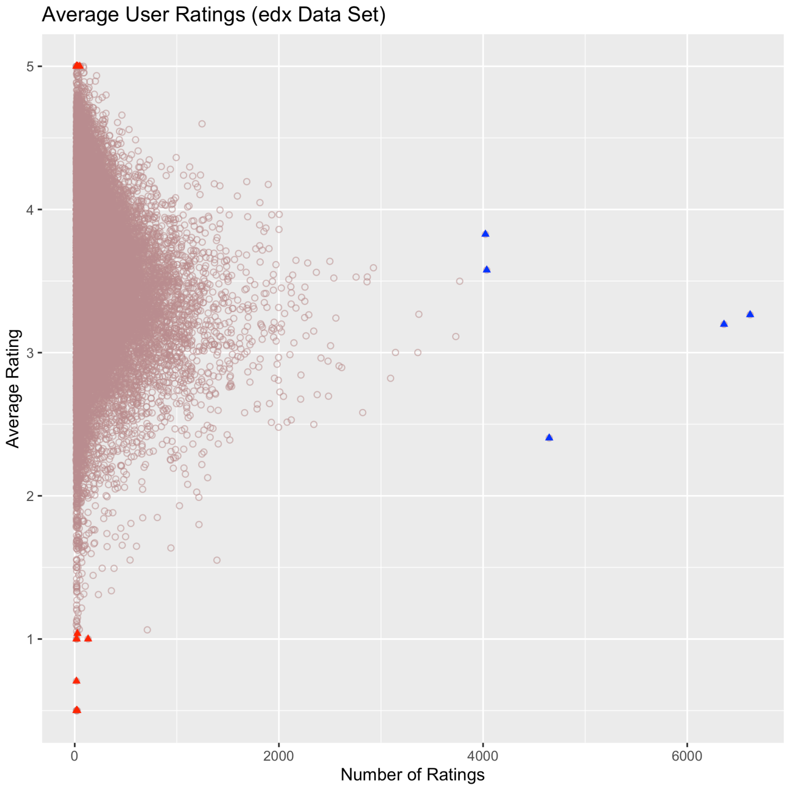

Figure 2: Average User Ratings

As for movies, the plot clearly show the highest and lowest raters by average rating are outliers. The pattern clearly shows how as the number of ratings increase, the average rating converges towards the centre. It's interesting that our most prolific users are quite different in their average rating, suggesting that they have a bias in one direction or the other from the mean user rating.

It's clear from the plot that the correlation between number of ratings and average rating is very weak. In fact it's a weak negative correlation (ρ = -0.1550551), suggesting users become slightly more critical as they rate more movies.

Time-dependent Variations

It is logical to think that the popularity of a movie changes over time. It might be a "hit" that starts off with good ratings when it is released, then as time goes on, opinions change as to how good it was. Or it might be a "slow burner" that improves with age. It's likely different movies go in and out of fashion over time.

The first plot in figure 3 shows how the average rating of our top 5 movies by the number of ratings has varied. It indicates a relatively long term variation in a movie's average rating over time.

When it comes to users who rate movies, then human nature comes into the mix. Rating then depends on all sorts of factors dependent on, by example, mood, fashion, or who they a watch with. The second plot in figure 3 shows how the ratings of the top 5 most prolific users vary over time. In indicates a lot of variability between users and a shorter term volatility in their ratings.

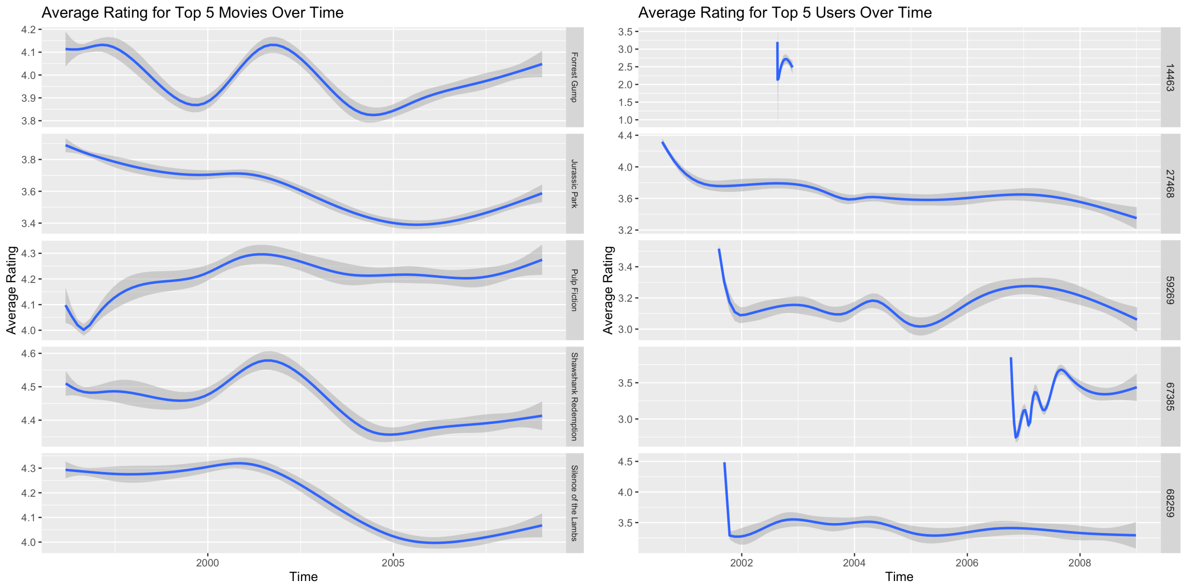

Figure 3: Time-Dependent Rating Variation

In modelling these two time-dependent biases, it makes sense to treat them separately with different time scales over which to measure the variation.