MovieLens_Report

Ian Gledhill

03/06/2019

Introduction

Overview

The web site http://www.movielens.org/ allows users to rate films from one to five stars, with half-star ratings allowed. This project aims to use the large number of movies, ratings and users available to find a method to predict movie ratings.

This report describes the process and methodology adopted by this project to analyze the existing movie ratings and to devise a method to predict ratings. The accuracy of the various methods will be compared, the accuarcy being indicated by the RMSE of the predictions of each method.

The project will analyze a dataset of movie ratings to predict what ratings a particular user would give a movie.

The users are randomly selected and are only identified by a uniquely assigned user id; there is no demographic and geopgraphic location information used.

Dataset

Source

The source dataset was created by the University of Minnesota from the online movie recommendation service MovieLens (https://movielens.org).

It is a large dataset and contains:

- 10000054 ratings
- 95580 tags
- 10681 movies by
- 71567 users

The source files can be obtained from this location:

- https://grouplens.org/datasets/movielens/10m/

This link shows further details, including usage rights provided by the license:

- http://files.grouplens.org/datasets/movielens/ml-10m-README.html

Sample Size for Analysis

This is a very large dataset and for the sake of managability a 10% sample of the source dataset will be used by the project. The source dataset is circa 1M ratings so the sample dataset will be approximately 100K ratings.

Shape of the Data

The source data contains three data sets:

- Movies
- Ratings
- Tags

Training and Test Data Partitions

The full approach is described later in this report but in summary the sample data is split into two partitions. A *training partition* will be allocated *80%* of the sample data and will be used to create (or *train*) the algorithm. A *test partition* will be allocated the remaining *20%* of the sample data and will be used to validate the predictions of the algorithm.

Goal

The objective of the project was to find a movie recommendation algorithm. Its success will be measured by reference to the RMSE produced when the algorithm is validated against the test dataset. The RMSE will be graded against a predfined set of thresholds to determine the overall quality of the algorithm (described later).

Approach

Data Cleansing

Modelling Techniques

The algoritm will start from a naive position and then be refined as further insights into the data are gained. The effect of each refinement will be accessed using a loss function. The RMSE (*residual mean squared error*) will be used as the loss function for this project.

Version 1 : Naive Implementation

The first attempt will use the average movie rating of the training data to predict the rating in the test data. The RMSE will then be measured. The formula for Y (the prediction) is described as follows, where u is the user, i is the movie, μ is the average and ϵ is the error.

$$Y_{u,i} = \mu + \epsilon_{u,i}$$

The result for this is shown here:

Table 1: Results for method 1

method	RMSE
Mean	1.060006

Version 2 : Consider the individual movie effect

This modification will use the mean with the addition of a movie effect to take into account that some movies are rated more favourably then others. For example, the top 5 most-rated and top 5 least-rated movies are shown below:

Table 2: Top 5 Movies - Best Rated

title	n	average
Shawshank Redemption, The (1994)	27988	4.456928
Godfather, The (1972)	17776	4.418851
Usual Suspects, The (1995)	21533	4.369967
Schindler's List (1993)	23234	4.365628
Sunset Blvd. (a.k.a. Sunset Boulevard) (1950)	2918	4.323852

Table 3: Top 5 Movies - Worst Rated

title	n	average
Horrors of Spider Island (Ein Toter Hing im Netz) (1960)	31	1.0967742
PokÃ©mon Heroes (2003)	145	1.0034483
Disaster Movie (2008)	31	0.9838710
From Justin to Kelly (2003)	198	0.9292929
SuperBabies: Baby Geniuses 2 (2004)	59	0.8135593

We can see that some movies are rated very highly. For example Shawshank Redemption, The (1994) has an average rating of 4.456928 stars whilst SuperBabies: Baby Geniuses 2 (2004) only has 0.8135593. The best and worst movies are not necessilly the most and least frequently rated. The top 5 most and least rated are shown below:

Table 4: Top 5 Movies - Most Number of Ratings

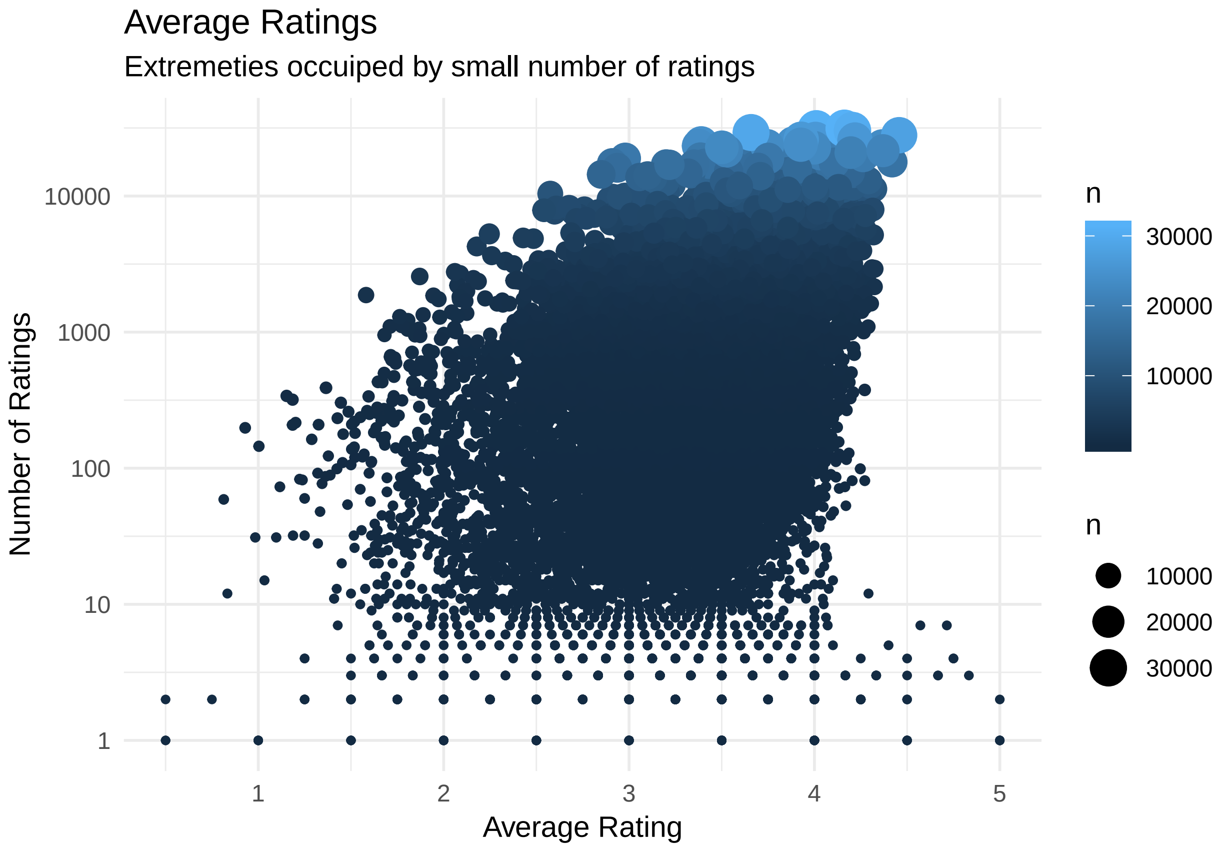
title	n	average
Pulp Fiction (1994)	31336	4.161731
Forrest Gump (1994)	31076	4.010265
Silence of the Lambs, The (1991)	30280	4.205086
Jurassic Park (1993)	29291	3.658189
Shawshank Redemption, The (1994)	27988	4.456928

Table 5: Top 5 Movies - Least Number of Ratings

title	n	average
Where A Good Man Goes (Joi gin a long) (1999)	1	4.0
Wings of Eagles, The (1957)	1	3.5
Women of the Night (Yoru no onnatachi) (1948)	1	4.0
Won't Anybody Listen? (2000)	1	2.0
Zona Zamfirova (2002)	1	4.0

We can see that some movies are rated very often. For example Pulp Fiction (1994) has 31336 ratings whilst Women, The (2008) only has 21.

If we plot the mean against the number of ratings we can see that the extremeties (1 star and 5 star) are only obtained by movies with a small number of ratings.



This shows us that movies with small number of ratings may have a distorting effect.

The formula for Y (the prediction) this is described as follows, where u is the user, i is the movie, μ is the average, ϵ is the error and b is the movie effect.

$$Y_{u,i} = \mu + b_i + \epsilon_{u,i}$$

Table 6: Results for method 1 and method 2

method	RMSE
Mean	1.0600060
Movie Effect Model	0.9427515

Version 3 : Consider the individual user effect

This is similar to version 2, but instead of looking at the movie effect we're going to look at the user effect; i.e. the fact that some users generally rate movies more generously then others and some users make lots of reviews whilst others make very few.

For example, the top 5 and bottom 5 users based on avarege rating given by that user is shown below:

Table 7: Top 5 Users - Highest Average Ratings

userId	n	average
1	21	5

userId	n	average
1686	21	5
7984	18	5
11884	19	5
13027	31	5

Table 8: Top 5 Users - Lowest Average Ratings

userId	n	average
63381	20	0.525
13496	19	0.500
48146	27	0.500
49862	18	0.500
62815	19	0.500

As with movies we can see that there is variability in both the ratings given (some users appear to be more generous) and the number of ratings the usersa provide. For example 1 has given average rating of 5 stars whilst 62815 has given an average rating of only 0.5. This might be because user tends to be make unlucky movie choices and watches poor movies.

Table 9: Top 5 Users - Most Number of Ratings Provided

userId	n	average
59269	6637	3.264653
67385	6376	3.196910
14463	4637	2.406081
68259	4056	3.560651
27468	4018	3.831011

Table 10: Top 5 Users - Least Number of Ratings Provided

userId	n	average
66247	14	3.214286
66573	14	3.500000
22325	13	3.923077
60448	13	4.000000
64070	13	3.269231

The formula for Y (the prediction) is described as follows, where u is the user, i is the movie, μ is the average, ϵ is the error and b is the movie effect.

$$Y_{u,i} = \mu + b_u + \epsilon_{u,i}$$

Table 11: Results for method 1, 2 and 3

method	RMSE
Mean	1.0600060
Movie Effect Model	0.9427515
User Effect Model	0.9782952

Version 4

This will use the combination of the movie effect and the user effect together with an adjustment (a *regulator*) to predict a user's rating.

The formula for Y (the prediction) is described as follows, where u is the user, i is the movie, μ is the average, ϵ is the error and b is the movie effect.

$$\frac{1}{N}\sum_{u,i}(y_{u,i}-\mu-b_i)^2+\lambda\sum_{i}b_i^2$$

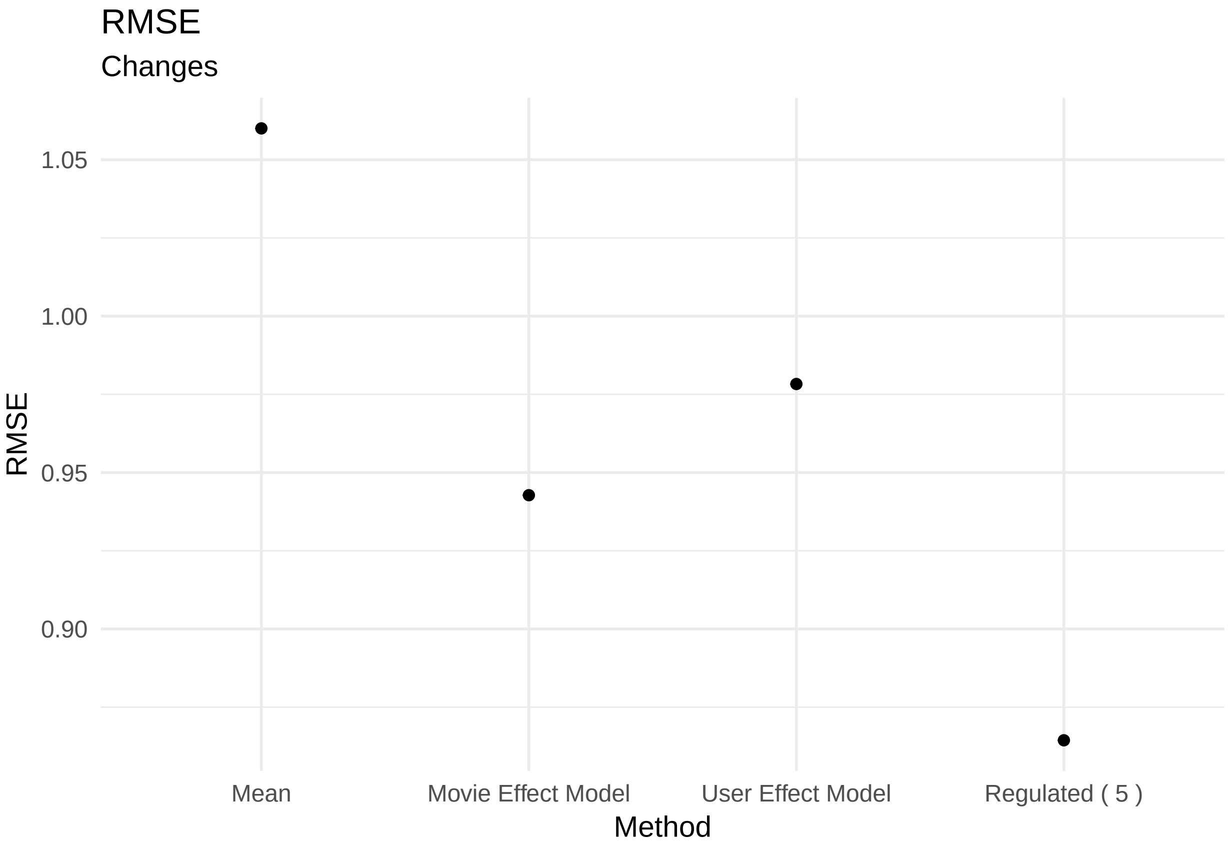
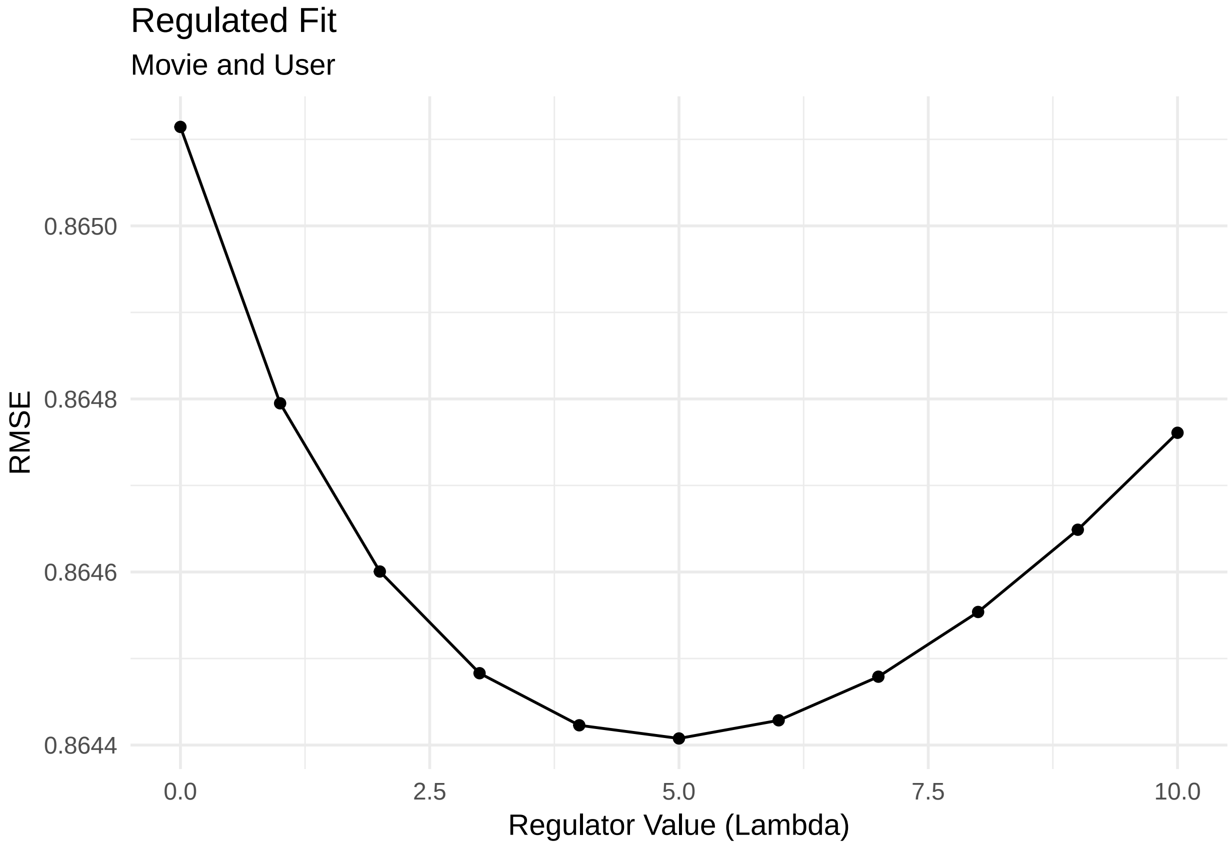


Table 12: Results for all 4 methods

method	RMSE
Mean	1.0600060
Movie Effect Model	0.9427515
User Effect Model	0.9782952

method	RMSE
Regulated (5)	0.8644076

Results


Regulated Fit
Movie and User
RMSE
0.8650
0.8648
0.8646
0.8644
0.0
2.5
5.0
7.5
10.0
Regulator Value (Lambda)

Conclusion

The best prediction occurs with the mean modified with the regulated movie and user effect. The majority of the improvement over using a simple mean comes when the user and movie effects are used in tandem. Using both is superior to using either by themselves. Also, the effect of the regulator value is not very significant compared to using just the movie and user effect.

Additional analysis that might increase the prediction accuracy (but not carried out by this project):

Are there additional effects which could be investigated: 1. genre : are some genres generally rated hire? 2. age : are older films generally rated higher or lower? 3. genre combinations : are some genre combinations more likely to rated higher or lower?

Bibliography and References

F. Maxwell Harper and Joseph A. Konstan. 2015. The MovieLens Datasets: History and Context. ACM Transactions on Interactive Intelligent Systems (TiiS) 5, 4, Article 19 (December 2015), 19 pages. DOI=http://dx.doi.org/10.1145/2827872