Analysis of MovieLens Dataset

Mark Athas

2022-08-25

Abstract

Companies in the streaming video service business, particularly those that provide access to television programs, blockbuster movies and classic films often include suggestive features labeled, "recommendations" or "recommended for you". These features rely on automated software to provide users with program recommendations suited to their individual interests, preferences and viewing experience. Such systems rely on recommendation models to identify movies to recommend. Thus, identification of the most optimal selection for each viewer is critical for a successful recommendation system. Using the *movielens* data as provided, multiple recommendation models will be considered. Various formulaic approaches will be applied to identify an algorithm that produces the most optimal model. These models will include both un-regularized and regularized approaches. Model effectiveness will be measured by the minimized Residual Mean Squared Error (RMSE).

Introduction

The source data for this project is *movielens* data prepared into a data set called edx. The edx data set contains approximately 9 million ratings of 10,677 movies which were recorded by 69,878 users. The edx data will be used as the training data set.

The single outcome from this project is to identify the optimal regularization model to apply to the data, and appropriate parameters that minimize the RMSE using the equation:

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

Where, i, u is the rating for movie i by user u, and $\hat{y}_{i,u}$ is the prediction. This RMSE equation will used for cross-validation on each model.

Also, supplied for this project is a validation data set that will be used as the test set in this project.

The approach to this project is to apply four models to the test data set provided. The first two models are un-regularized the last two models are regularized. Four steps will be applied in this analysis. The first step will provide a general description of the data and its structure. Next, the models will be described and applied to the train data with an explanation of each outcome. Third, conclusions will be drawn about the optimized model and what is says about the data. Finally, recommendations for further analysis will be offered.

Methodology

The first step in the analysis of the edx data is to understand its structure. The edx data is comprised of 6 columns which have names and types as listed in Table 1.

Table 1: Structure of edx Data Set

Column	Type
userId	integer
movieId	numeric
rating	numeric
timestamp	integer
title	character
genres	character

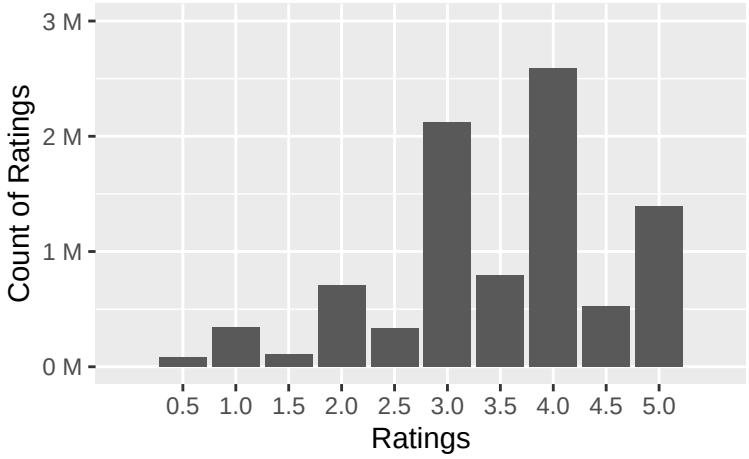
The 9 million ratings in edx have an average rating of 3.5 and standard deviation of 1.06. These and additional descriptive statistics are listed in Table 2.

Table 2: Descriptive Metrics of edx Data Set

Measure	Value
Number of Ratings	9,000,055
Number of Movies Rated	10,677
Number of Users Providing Ratings	69,878
Average Rating	3.512
Standard Deviation of Rating	1.06033

The distribution of the ratings is shown in Figure 1. It is demonstrative of a data set where ratings tend left skewed. Thus, a significant majority of the ratings are 3.0 or above. It interesting that the whole number and fractional number ratings have a similar skewness pattern, where the first four values increase relative to each other and the fifth falls between the 3^{rd} and 4^{th}. Because of rating skewness, 82.4 percent of the ratings are three or larger.

Figure 1: Total User Ratings by Rating



Rational for Regularization

Stratification of ratings into three tiers is included in Table 3. It reflects movies stratified by the number of ratings. Movies with the lowest number of ratings have an average rating lower than all movies while the movies with the highest number of ratings have an average rating higher than all movies This suggests there may be outliers in both extremes that may be mediated by regularization.

Table 3: Stratified Key Measures

Stratum	n	Mean	SD	Skew
Lowest 30 Rated Movies	36,956	2.9964	1.1136	-0.0098
All Ratings	9,000,055	3.5125	1.0603	-1.3794
Top 30 Rated Movies	707,860	3.9337	0.9617	-0.2069

Analysis

The analysis will apply four models to the edx data set treating it as a train data set. A second provided data set is a validation data set which will be treated as a test data set. The models are identified as M1 through M4 and have the modeling intent as described in Table 4.

Table 4: Models Used in this Analysis

Method	Formula
Just the average	$Y_{u,i} = \mu + \epsilon_{u,i}$
M1: Movie Effect Model	$Y_{u,i} = \mu + b_i + \epsilon_{u,i}$
M2: Movie + User Effects Model	$Y_{u,i} = \mu + b_i + b_u + \epsilon_{u,i}$
M3: Regularized Movie Effect Model	$\sum_{u,i}(y_{u,i} - \mu - b_i)^2 + \lambda \sum_i b_i^2$
M4: Regularized Movie + User Effect Model	$\sum_{u,i}(y_{u,i} - \mu - b_i - b_u)^2 + \lambda(\sum_i b_i^2 + \sum_u b_u^2)$

Just the Average

The first data presented is not an evaluated model but is the simply the RMSE of the train data against the means. This is done by summing the squared difference between each rating in the training set and the mean, and then taking the square root of that sum, as shown in the RMSE formula in the Introduction. For *Just the Average*, no prediction is made using the validation data set, and thus no cross-validation. The RMSE is 1.0612. This is provided for reference and comparison to the actual models as it presumed that the models will perform better than the average.

Model 1: Movie Effect Model (M1)

This model employs cross-validation using the provided validation data as test data and is labeled the *M1: Movie Effect Model*. This model recognizes that some movies are ordinarily rated higher than others. It extends the *Just the Average* by adding a bias variable b_i to represent the average difference between each movie rating and the mean of the of all ratings for that movie using the test data set. The formula is $b_i = \frac{1}{N}\sum_i m_i - \mu$. The value of b_i is then added to the mean to form a predicted rating for the corresponding movie i. The RMSE is then calculated as described before. The *M1: Movie Effect Model* produces an RMSE of 0.94391.

The M1 model does improve RMSE, but it is still lacking in recognizing the 'best' movies. Table 5 lists the ten highest rated movies per the M1 model. It is evident that these are not widely distributed and well known movies. Therefore, our predictions for top movies for that highly selective movie watcher could get recommendations like those listed.

Table 5: Top Rate Movies in edx Data Set (Un-regularized)

Title	Average_Rating	n
Hellhounds on My Trail (1999)	5.0	1
Satan's Tango (Sátántangó) (1994)	5.0	2
Shadows of Forgotten Ancestors (1964)	5.0	1

Title	Average_Rating	n
Fighting Elegy (Kenka erejii) (1966)	5.0	1
Sun Alley (Sonnenallee) (1999)	5.0	1
Blue Light, The (Das Blaue Licht) (1932)	5.0	1
Who's Singin' Over There? (a.k.a. Who Sings Over There) (Ko to tamo peva) (1980)	4.8	4
Human Condition II, The (Ningen no joken II) (1959)	4.8	4
Human Condition III, The (Ningen no joken III) (1961)	4.8	4
Constantine's Sword (2007)	4.8	2

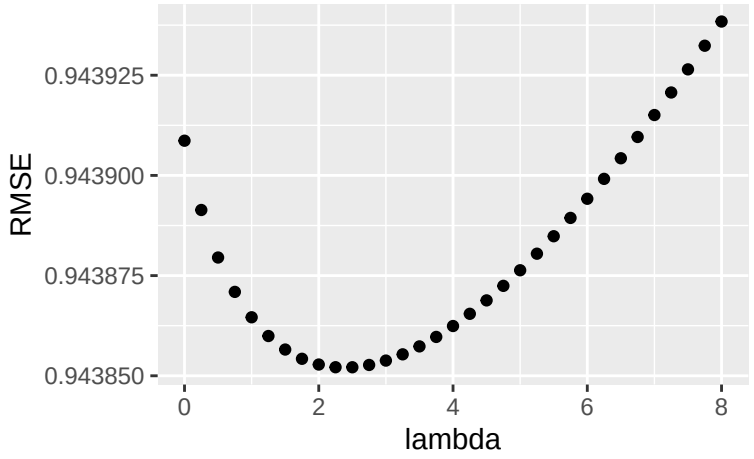
Model 2: Movie + User Effect Model (M2)

The *M2: Movie + User Effect Model* is also cross-validated using the validation data. This model recognizes that not only movies have biased ratings, but the users who rate those movies hold a bias based on their personal preferences. The M2 model will add an additional user bias variable b_u to the M1 model to lessen the effects of both movie and user bias. The user bias variable is similar to M1 bias variable as it represents the average difference between each user rating and the overall mean less the previously calculated movie bias b_i. The formula for user bias rating is $b_u = \frac{1}{N}\sum_i u_i - \mu - b_i$. The RMSE is then calculated as described before. The *M2: Movie + User Effect Model* produces an RMSE of 0.86535.

Model 3: Regularized Movie Effect Model (M3)

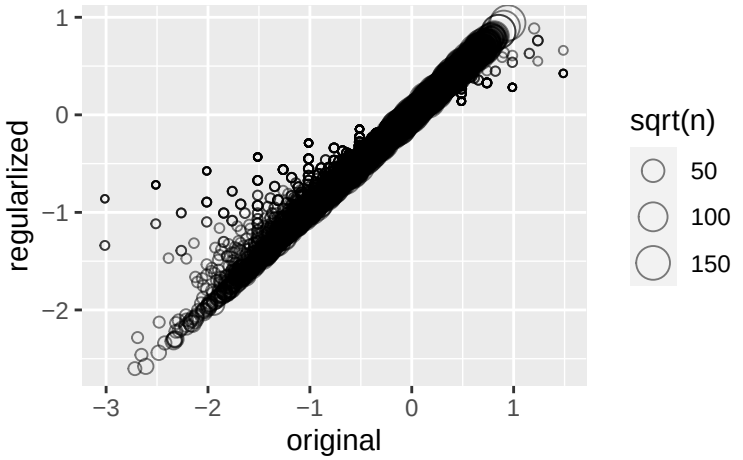
This model is also tested against the validation data. It includes the b_i and u_i but adds an additional quantity lambda l in the bias calculation. The resulting bias movie formula becomes $b_i = \frac{1}{N+l}\sum_i m_i - \mu$. To find best value for l the formula is applied over a range of 0 to 8, with a stepping interval of 0.25. Figure 2 shows the results of each evaluated lambda. The value of l that produces the smallest RMSE is 2.5.

Figure 2: Determination of Optimal l for M3 Model



The effect of lambda is to pull the extreme ratings values (near 0 or near 5) of movies with few raters toward the mean of all ratings. To visualize this pull Figure 3 shows the original data on the x-axis (from the M1 Model), and the regularized value on the y-axis (from this M3 model). Movies with a low number of ratings are pulled further toward the mean, while movies with a large number of ratings are pulled little or not at all. The RMSE is then calculated as described before. The *M3: Regularized Movie Effect Model* produces an RMSE of 0.94385. Note that this is 'worse' than *M2: Movie + User Effect Model*, which includes user bias, but better than the movie only bias in the *M1: Movie Effect Model*.

Figure 3: Impact of Regularization using M3 Model



With the movie bias effect remediated, Table 6 shows the new list of top rated movies. This is a more relevant list of top and well viewed movies as compared with Table 5 above.

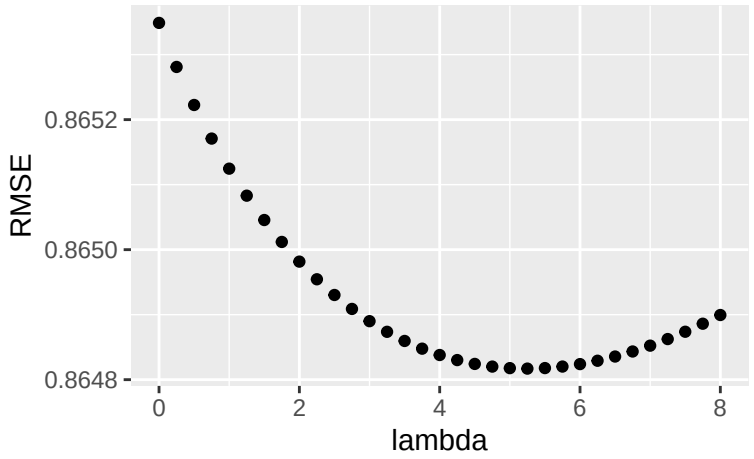
Table 6: Top Rate Movies in edx Data Set (Regularized) _

Title	Average_Rating	n
Shawshank Redemption, The (1994)	4.5	28,015
Godfather, The (1972)	4.4	17,747
More (1998)	4.7	7
Usual Suspects, The (1995)	4.4	21,648
Schindler's List (1993)	4.4	23,193
Casablanca (1942)	4.3	11,232
Rear Window (1954)	4.3	7,935
Sunset Blvd. (a.k.a. Sunset Boulevard) (1950)	4.3	2,922
Third Man, The (1949)	4.3	2,967
Double Indemnity (1944)	4.3	2,154

Model 4: Regularized Movie + User Effects Model (M4)

This final model is also uses cross-validation with the validation data. As with M3, a lambda is added to the equation $b_u = \frac{1}{N+l}\sum_i u_i - b_i - \mu$. And again, the formula was run with lambda over a range of 0 to 8, with a stepping interval of 0.25 (see Figure 4). The value of l that produces the smallest RMSE is 5.25. The result of M4 is to regularize the extreme positive (5) and negative (0) user ratings for movies that have few raters. The RMSE is then calculated as described before. The *M4:Regularized Movie + User Effects Model* model produces an RMSE of 0.86482.

Figure 4: Determination of Optimal l for M4 Model



Results

A summary of the results from all four models is provided in Table 7.

Table 7: Summary of All Methods

Method	RMSE
Just the average	1.0612018
M1: Movie Effect Model	0.9439087
M2: Movie + User Effects Model	0.8653488
M3: Regularized Movie Effect Model	0.9438521
M4: Regularized Movie + User Effects Model	0.8648170

Three patterns of model performance improvement emerge from Table 7. First, is the progression of *Just the Average* to *M1: Movie Effect Model* to *M2: Movie + User Effects Model*. Models M1 and M2 seek to address the extreme positive and negative influence of raters that demonstrate a love/hate tendency toward movies. The improvement is striking as these model show a step change of -0.11729 and -0.07856 respectively. Total improvement over *Just the Average* is -0.19585 which is a 18.5 percent change. This improvement in model performance is the result of the movie bias reduction from b_i as applied to movie ratings, and user bias reduction b_u applied to user ratings.

The Second pattern of improvement in Table 7 is in the chain of *M1: Movie Effect Model* to *M3: Regularized Movie Effect Model*. Here, M3 adds regularization to further address movies that have a relatively low number of ratings. The improvement of M3 over M1, is slight at -5.65e-05, with a percent change of 0.01.

The Final pattern of improvement in Table 7 is in the chain of *M2: Movie + User Effects Model* to *M4: Regularized Movie + User Effects Model*. Here, M4 adds regularization to address movie and users that have a relatively low number of ratings. The improvement of M4 over M2, is slight at -0.00053, with a percent change of -0.06. This is also a slight improvement, yet still is better model.

Conclusion

This analysis evaluated four models, two un-regularized and two regularized. The un-regularized models sought to address movie and user rater bias across all movies. The regularized models addressed extreme movie ratings for those movies that had few ratings. All four models showed improvement over the *Just the Average* only with the *M4: Regularized Movie + User Effects Model* demonstrating the most improvement with a **RMSE of 0.86482**.

Additional research to consider includes a model that determines bias in movie series such as *Godfather I, II and III* as well as *Star Wars III-VIII*. These movies are not only popular and well watched, but are lauded by many. This may cause a tendency to overrate all movies in the series, and could be an opportunity to analyze the data and models further. The edx data set does currently include date and time data that could provide opportunity for seasonal and time of day analysis. Also, a genre effect could exist in the data resulting in families of movies that have differential ratings. Finally, if user geographic data was available an analysis for geographic regions differentiation could be modeled.