MovieLens Recommendation System Capstone Project

Dave Downing

5/1/2020

Contents

1 Introduction

This is a report on optimizing a movie recommendation system to provide more accurate predictions based on historical rating data. To complete this analysis, a subset of the MovieLens data set obtained from *https://grouplens.org/datasets/movielens/10m/* was used. The data set includes over 9,000,000 movie reviews on over 10,000 different movies.

The simplest way to predict a rating for a particular movie from a particular user would be to simply use the average (mean) of all ratings from all users and use this average to predict all ratings. The below shows the mean as well as the residual mean squared error (RSME) from this data set:

	x
Mean	3.512465
RMSE	1.061202

The goal of this project was to apply machine learning techniques to lower the RMSE to below 0.86490 which was able to be achieved by examining the following:
1. The Movie Effect - Understanding that some movies score better or score worse than others
2. The User Effect - Understanding that some users give higher or lower scores than others
3. The Regularization Effect - Understanding that more uncertainty due to the smaller sample size impacts both:
i) Movies that have received a very small number of reviews
ii) Users that have given a very small number of reviews

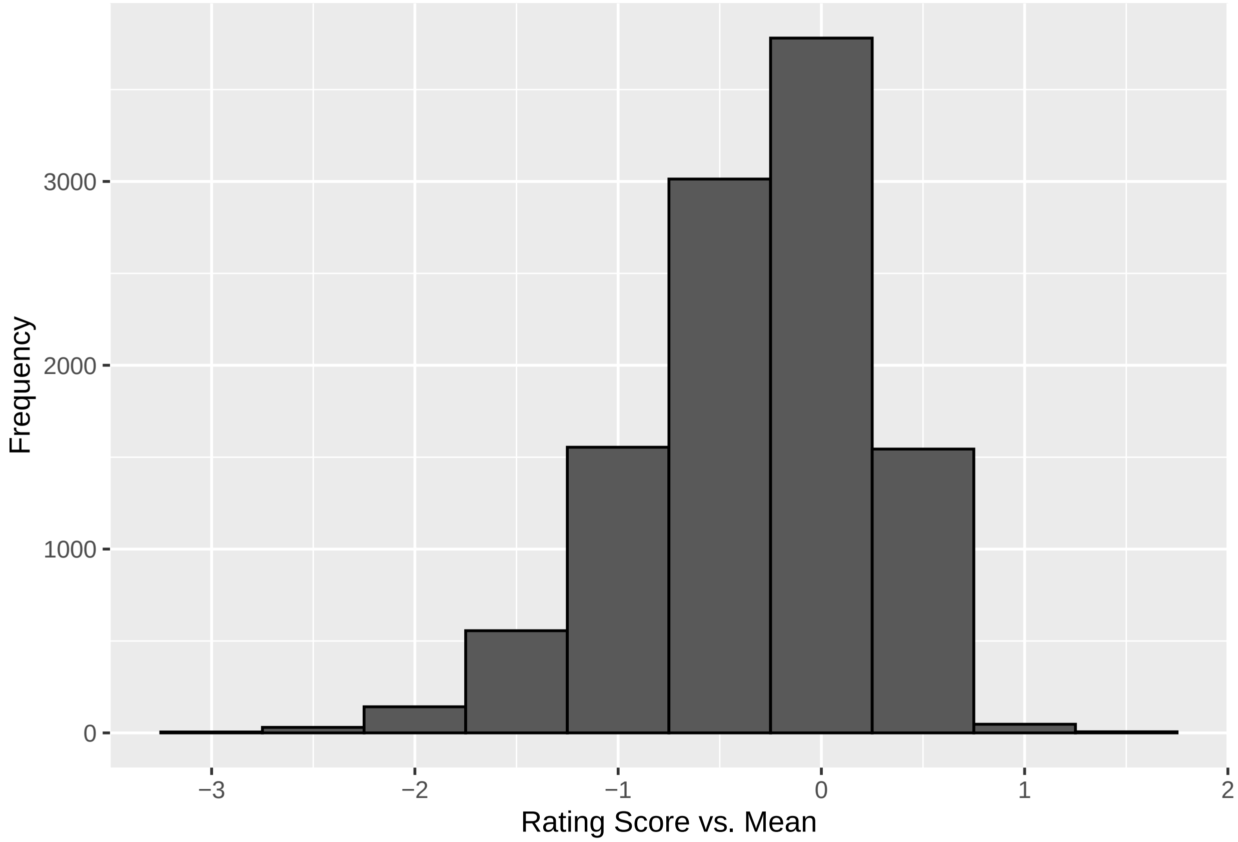
By controlling for these factors the RMSE goal was achieved with a final RMSE of **0.8648**.

2 Methods / Analysis

In order to avoid over-fitting our model, the data was split into a training set and a test set. Only the training set was used to create the models. The test set was only used for validation of the model results as a control for the RMSE.

2.1 Movie Effect

To begin, an examination was taken of the effect caused by the movies themselves - simply that some movies are better than others. This can clearly be seen below in that some movies score much higher than others and vice versa:



This movie effect an be accounted for by using the following equation:

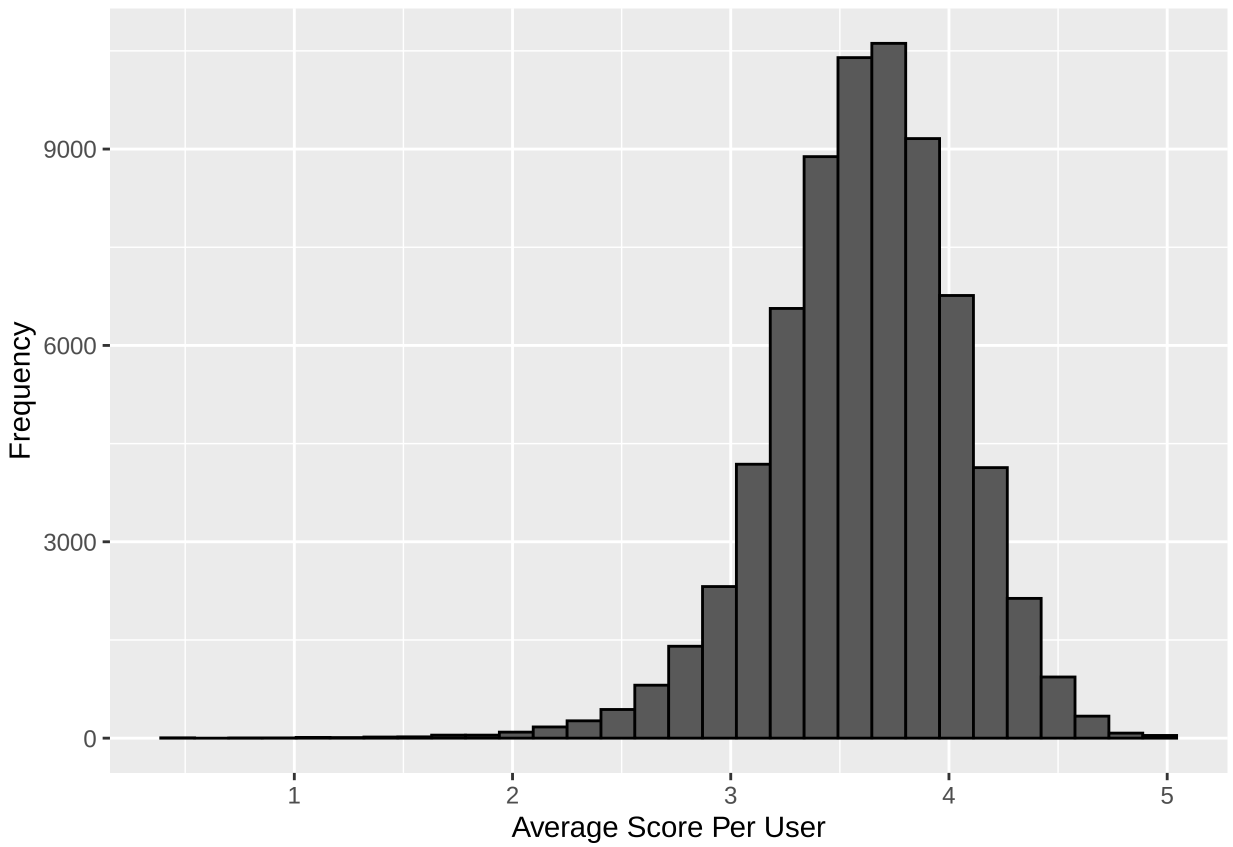
$$Y_{u,i} = \hat{\mu} + b_i + \epsilon_{u,i}$$

Where $\hat{\mu}$ is the mean, $\varepsilon_{i,u}$ is the independent errors sampled from the distribution centered at 0 and the term b_i is used for bias control of movie i as a measure of popularity.

This effect can be applied to our model and denoted as the **Movie Effect Model**.

2.2 Movie + User Effect

The next step was to examine the user effect. Looking at all users who had at least 50 reviews, the same trend that was just illustrated for movies also applies for users - some users are much harsher with their ratings and some are much more generous:



To control for this a new term b_u can be added to the equation so that the formula now becomes:

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

The term b_u is added to measure for the bias of user u. This effect can be applied to our model and denoted as the **Movie + User Effect Model**.

2.3 Regularized Movie Effect

Looking at both the predicted best and predicted worst movies from the model to this point, something very interesting can be noticed:

Best Predicted Movies:

title	b_i	n
Hellhounds on My Trail (1999)	1.487535	1
Satan's Tango (SÃ¡tÃ¡ntangÃ³) (1994)	1.487535	2
Shadows of Forgotten Ancestors (1964)	1.487535	1
Fighting Elegy (Kenka erejii) (1966)	1.487535	1
Sun Alley (Sonnenallee) (1999)	1.487535	1
Blue Light, The (Das Blaue Licht) (1932)	1.487535	1
Who's Singin' Over There? (a.k.a. Who Sings Over There) (Ko to tamo peva) (1980)	1.237535	4
Human Condition II, The (Ningen no joken II) (1959)	1.237535	4
Human Condition III, The (Ningen no joken III) (1961)	1.237535	4
Constantine's Sword (2007)	1.237535	2

Worst Predicted Movies:

title	b_i	n
Besotted (2001)	-3.012465	2
Hi-Line, The (1999)	-3.012465	1
Accused (Anklaget) (2005)	-3.012465	1
Confessions of a Superhero (2007)	-3.012465	1
War of the Worlds 2: The Next Wave (2008)	-3.012465	2
SuperBabies: Baby Geniuses 2 (2004)	-2.717822	56
Hip Hop Witch, Da (2000)	-2.691037	14
Disaster Movie (2008)	-2.653090	32
From Justin to Kelly (2003)	-2.610455	199
Criminals (1996)	-2.512465	2

Look how obscure most of these movies are. To compensate for this, regularization can be used to handle movies with just a small number of reviews. This method adds a component λ (lambda) that penalizes movies that are increasing the RMSE due to a small sample size. The following equation is used to optimize b_i to account for this sample size variance:

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

This effect can now be added to the model and denoted as the **Regularized Movie Effect Model**. After making this adjustment, take a look at the best and worst list now:

Updated Best Predicted Movies:

title	b_i	n
Shawshank Redemption, The (1994)	0.9425819	28015
Godfather, The (1972)	0.9027736	17747
More (1998)	0.8855520	7
Usual Suspects, The (1995)	0.8532899	21648
Schindler's List (1993)	0.8509364	23193
Casablanca (1942)	0.8077788	11232
Rear Window (1954)	0.8059324	7935
Sunset Blvd. (a.k.a. Sunset Boulevard) (1950)	0.8027275	2922
Third Man, The (1949)	0.7982878	2967
Double Indemnity (1944)	0.7974264	2154

Updated Worst Predicted Movies:

title	b_i	n
SuperBabies: Baby Geniuses 2 (2004)	-2.601676	56
From Justin to Kelly (2003)	-2.578067	199
Disaster Movie (2008)	-2.460837	32
PokÃ©mon Heroes (2003)	-2.438765	137
Carnosaur 3: Primal Species (1996)	-2.338264	68
Glitter (2001)	-2.319841	339
Pokemon 4 Ever (a.k.a. PokÃ©mon 4: The Movie) (2002)	-2.305711	202
Gigli (2003)	-2.300797	313
Barney's Great Adventure (1998)	-2.297353	208
Hip Hop Witch, Da (2000)	-2.283304	14

Now that makes a lot more sense as the best list is now littered with movies widely considered classics. Unfortunately for fans of *Superbabies: Baby Geniuses 2* and *From Justin to Kelly* - not only do they still make the worst list but they are now in the top 2 spots.

2.4 Regularized Movie + User Effect

In addition to these movies with a very low number of reviews, there are also users that have a very small number of reviews which also has an impact on the RMSE. So the model needs to be adjusted to handle small sample sizes for users in addition to the movies. The exact same regularization method that was just applied to the movie bias can be applied to the user bias. With these improvements the model can be updated and denoted as **Regularized Movie + User Effect Model**.

3 Results

Here are the results of the RSME for each of the denoted models:

method	RMSE
Simple Average	1.0612018
Movie Effect Model	0.9439087
Movie + User Effects Model	0.8653488
Regularized Movie Effect Model	0.9438521
Regularized Movie + User Effect Model	0.8648170

The above RMSE table shows that once the movie effect is introduced, some improvement in accuracy is achieved. However, once the model is adjusted to account for both the movie AND the user effect, much better improvement is seen. In fact, the goal RMSE is almost achieved with this method. However, while close to the objective, this model is not quite there just yet. Next by adding the regularization effect to control for movies with small numbers of reviews, the RMSE is improved a bit vs. using the movie effect model without regularization. Once the model puts it all together and adds in regularization to also account for users with smaller number of reviews, the goal of getting below the target RMSE of 0.8649 is achieved with a final RSME of **0.8648**.

4 Conclusion

By applying these methods, a very significant improvement to the predicted movie ratings was achieved. There is however a very significant factor that has not been addressed in this analysis. An important source of variation comes from the fact that certain groups of movies and certain groups of users have very similar rating patterns. It is possible that these patterns could be observed by studying the residuals and converting the data into a matrix where each user gets a row and each movie gets a column. Matrix factorization could then be performed to see if the model can be improved upon even further.