Recommendations with Movielens Data

Jeff Warchall

11/13/2020

Executive Summary

The purpose of this analysis is to use the Movielens dataset to develop a recommendation system which predicts as accurately as possible the rating that a particular user will assign to a particular movie; specifically, a root mean square error between the predicted rating and the actual rating of less than **0.86490**. The Movielens data set used for this project contains approximately 10 million records. Each record represents a unique conbination of user and movie and contains the following observations for each record:

- User ID
- Movie ID
- Rating of this movie by this user out of 5 stars
- Timestamp of when the rating was made
- Movie Title
- List of Genres that the movie fits into

To perform this analysis the Movielens set was split into an approximately 90% / 10% partition with the larger set used to train a linear model and the smaller set used to validate the performance of the model. The linear model fit to the training set considered relationships between each of the observations and the rating with the three most significant effects being retained. The final result is a model which predicts the rating of a new unique movie and user combination using bias drawn from the user's past behavior in assigning ratings and bias effects from the particular movie's previous ranking and the ranking performance of the genre as a whole.

Methods and Analysis

A noteable feature of the dataset is that the release year of every movie is included parenthetically after the movie titles, but is not captured in its own field. Therefore, the first step was to extract this information and examine the effect that year of release might have on the average rating:

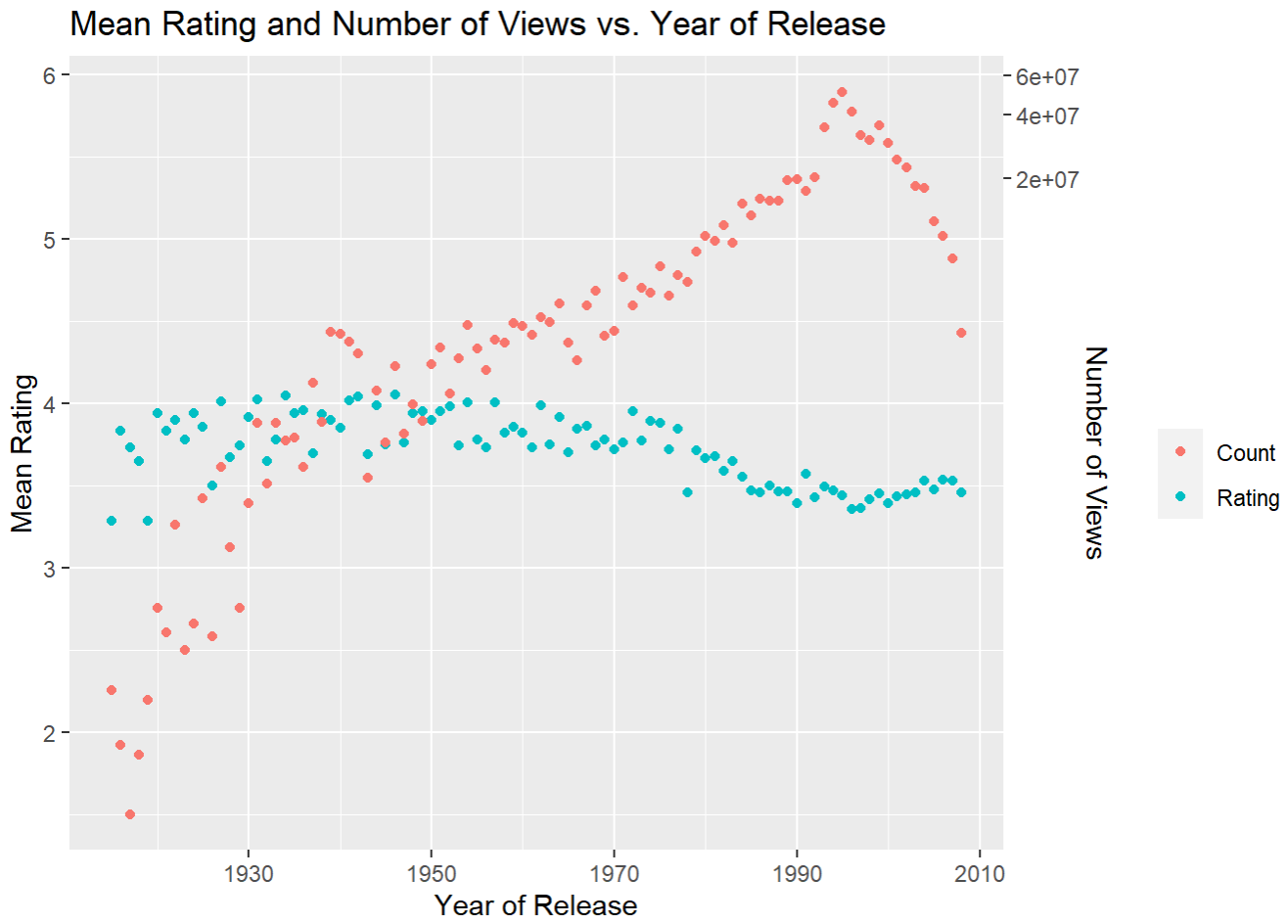

Figure 1: Release Year

There is a noticable decline in the mean rating of movies released after approximately 1970; however, there is also a clear increase in the views begining in approximately 1950. While this could indicate a strong effect where movies released after 1970 are generally worse than those released before 1970, it is also important to note that all of the ratings were given in the 1990's and later. This indicates that it is more likely that the older movies that are still being viewed and ranked today are likely to be some of the better movies overall. Thus, it is not that older movies are better, but rather that among older movies, it is only the good ones that are strongly represented in the data set.

Furthermore, just as some movies are infrequently represented in the data there are some users and genres that appear ony rarely in our training set:

title	count
1, 2, 3, Sun (Un, deuz, trois, soleil) (1993)	1
100 Feet (2008)	1
4 (2005)	1
Accused (Anklaget) (2005)	1
Ace of Hearts (2008)	1
Ace of Hearts, The (1921)	1

userId	count

userId	count
62516	10
22170	12
15719	13
50608	13
901	14
1833	14

genres	count
Action|Animation|Comedy|Horror	2
Action|War|Western	2
Adventure|Fantasy|Film-Noir|Mystery|Sci-Fi	2
Adventure|Mystery	2
Crime|Drama|Horror|Sci-Fi	2
Documentary|Romance	2

To account for under represented movies, users, or genres data regularization was attempted to assign less weight to bias for factors that occur only rarely in the data. Plotting the errors obtained with different values of lambda yields the following:

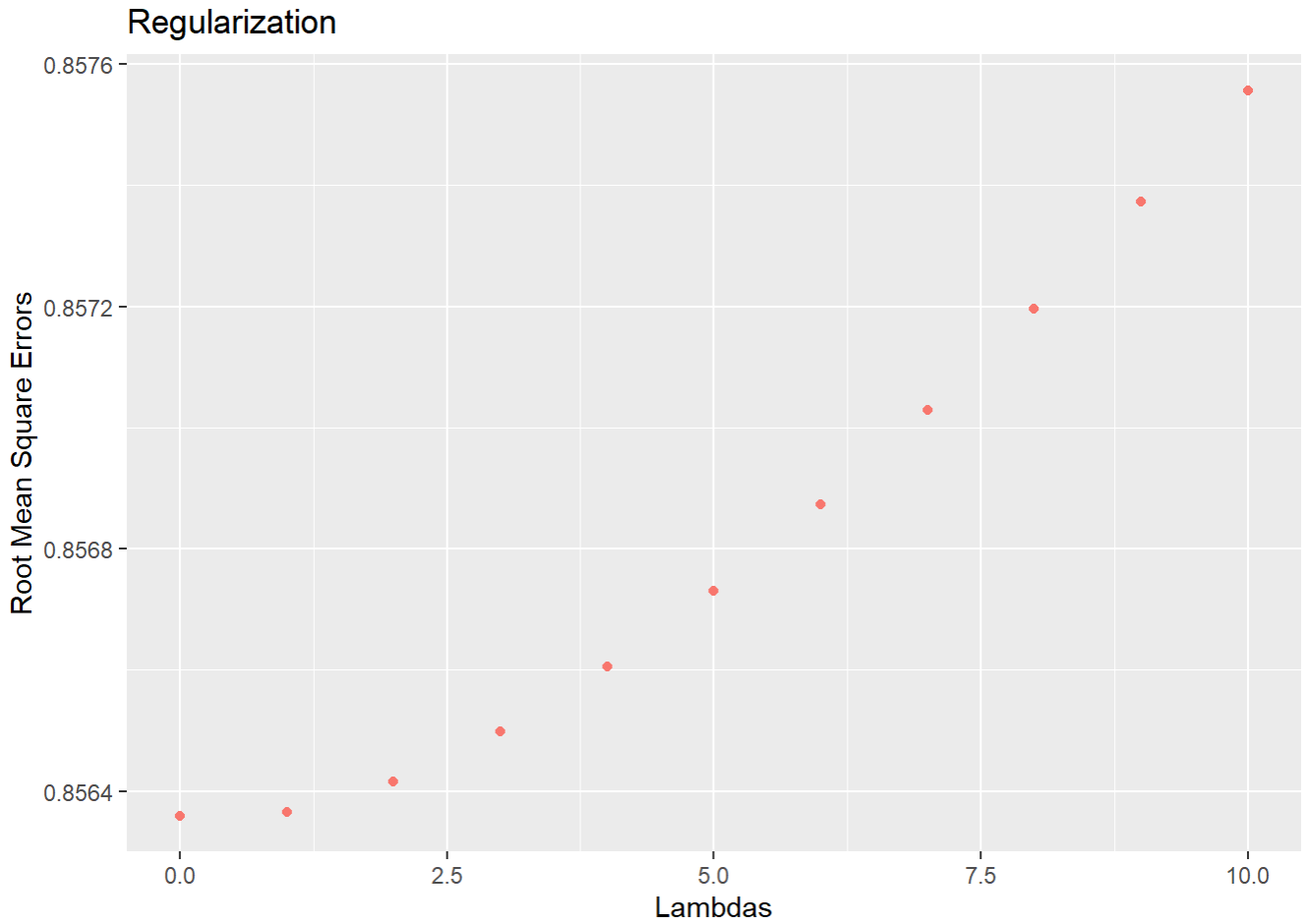

Figure 2: Selecting Lambda

Here it is seen that the least error can be achieved with a lambda of zero. Hence, although there are some instances of under represented observations, the effect on the final model is not substantial enough to obtain more accurate predictions through regularization. Therefore, the linear model for each effect does not includ a lambda value to tune the model.

The RMSE and within_bounds functions defined below are used to calculate the errors in our models and to restrict the predictions to the range 0.5 to 5 stars. For example, the model could predict that a good movie, rated by a generous user, in a highly rated genre could result in a predicted rating of more than 5 stars; the within_bounds function would reduce this prediction to 5.0 stars.

```
# This is used to restrict the predictions to within the minimum and maximum ratings that are po
ssible
# in the data set and slightly reduce errors, since the linear model developed can yield ratings
of greater
# than 5 and less than 0.5.
within_bounds <- function(x) {
  if(x > 5){
    result <- 5
  } else if (x < 0.5) {
    result <- 0.5
  } else {
    result <- x
  }
  return(result)
}

# This is used to calculate the rmse of the predictions
RMSE <- function(x,y){
  SE <- (x - y) ^ 2
  result <- mean(SE) ^ (1/2)
  return(result)
}
```

Based on the insights gained thus far, the prediction system was built using a linear model without regularization which considers genre, movie, and user bias. The accuracy of these three models applied individually to the training set are compared below, with the accuracy of merely predicting the mean rating in all cases included for comparison:

Methods	**RMSE**
Just the Average	1.0603313
Genre Bias	1.0179838
User Bias	0.9700086
Movie Bias	0.9423475

None of these factors are sufficient to achieve the desired accuracy alone. However, by combining the models to include all three factors an acceptable result can be obtained.

Results

Using the knowledge gained above, the final linear model is constructed as follows:

```
# calculate the average rating of the training dataset
mu <- mean(edx$rating)

# calculate the movie bias (b_m)
b_m <- edx %>%
  group_by(movieId) %>%
  summarize(b_m = mean(rating - mu))

# calculate the user bias (b_u)
b_u <- edx %>%
  left_join(b_m, by = "movieId") %>%
  group_by(userId) %>%
  summarize(b_u = mean(rating - mu - b_m))

#calculate the genre bias (b_g)
b_g <- edx %>%
  left_join(b_m, by = "movieId") %>%
  left_join(b_u, by = "userId") %>%
  group_by(genres) %>%
  summarize(b_g = mean(rating - b_m - b_u - mu))

# use the bias effects calculated above to generate predicted ratings on the training set.  Note
the use of the
# within_bounds function to restrict values to the range 0.5 to 5.0
train_prediction <- edx %>%
  left_join(b_m, by = "movieId") %>%
  left_join(b_u, by = "userId") %>%
  left_join(b_g, by = "genres") %>%
  mutate(pred = sapply(mu + b_m + b_u + b_g, within_bounds)) %>%
  .$pred
```

Using the linear model thus defined, the final result when the model is applied to the validation set is as follows:

```
## Validation Set RMSE =  0.8647516
```

This result is below the desired target and the model satisfies the requirements. However, there are significant errors in extreme cases, such as when a particular combination of genres is only shared by a small number of movies or when a particular user rated a small number of movies. This effect in the extreme cases might have been reduced using regularization, but as demonstrated above, the overall accuracy of the model was not improved with regularization.

Genres

genres	mean_error	n
Crime|Film-Noir|Romance	3.080747	1
Adventure|Comedy|Fantasy|Romance	2.037832	1
Comedy|Crime|Drama|Horror|Mystery	1.696922	3
Animation|Fantasy|Sci-Fi|War	1.685476	3
Drama|Fantasy|Horror|Sci-Fi	1.670714	3

genres	mean_error	n
Children|Comedy|Drama|Mystery	1.550403	3

genres	mean_error	n
Drama|Horror|Mystery|Sci-Fi|Thriller	0.0450405	1
Action|Crime|Drama|War	0.0781142	1
Action|Crime|Film-Noir	0.1032577	2
Film-Noir|Horror	0.1120558	2
Documentary|Romance	0.1256344	1
Adventure|Horror|Mystery|Thriller	0.1258603	3

Movies

title	mean_error	n
Uncle Nino (2003)	3.784550	1
They Live by Night (1948)	3.080747	1
Attack Force Z (a.k.a. The Z Men) (Z-tzu te kung tui) (1982)	3.037832	1
Tarantella (1995)	2.968180	1
It (1927)	2.923087	1
Nest, The (Nid de GuÃªpes) (2002)	2.900114	1

title	mean_error	n
Headless Body in Topless Bar (1995)	0.0000000	1
Parallel Sons (1995)	0.0000000	2
Winter Kills (1979)	0.0029804	1
Dagon (2001)	0.0065684	1
Jindabyne (2006)	0.0068644	1
Grill Point (Halbe Treppe) (2002)	0.0083476	1

Users

userId	mean_error	n
12217	3.783381	1
23333	3.669086	1
15162	3.649361	1

userId	mean_error	n
43789	3.623291	2
1007	3.537085	1
16550	3.483075	1

userId	mean_error	n
9333	0	2
10754	0	1
13951	0	1
14849	0	1
17792	0	2
17941	0	2

Conclusions and Future Work

Overall the model performed to the desired degree of accuracy. However, there is room for improvement. For example, the following figure shows that although the predicted ratings were centered near the mean of the actual ratings, the model often predicted a rating close to 3.5 stars whereas an actual 3.5 star rating was rather uncommon as compared to 3 or 4 star ratings. It may be possible to improve the accuracy of this model by introducing an additional bias against half-star ratings, or perhaps simply rounding intermediate predictions to the nearest full star.

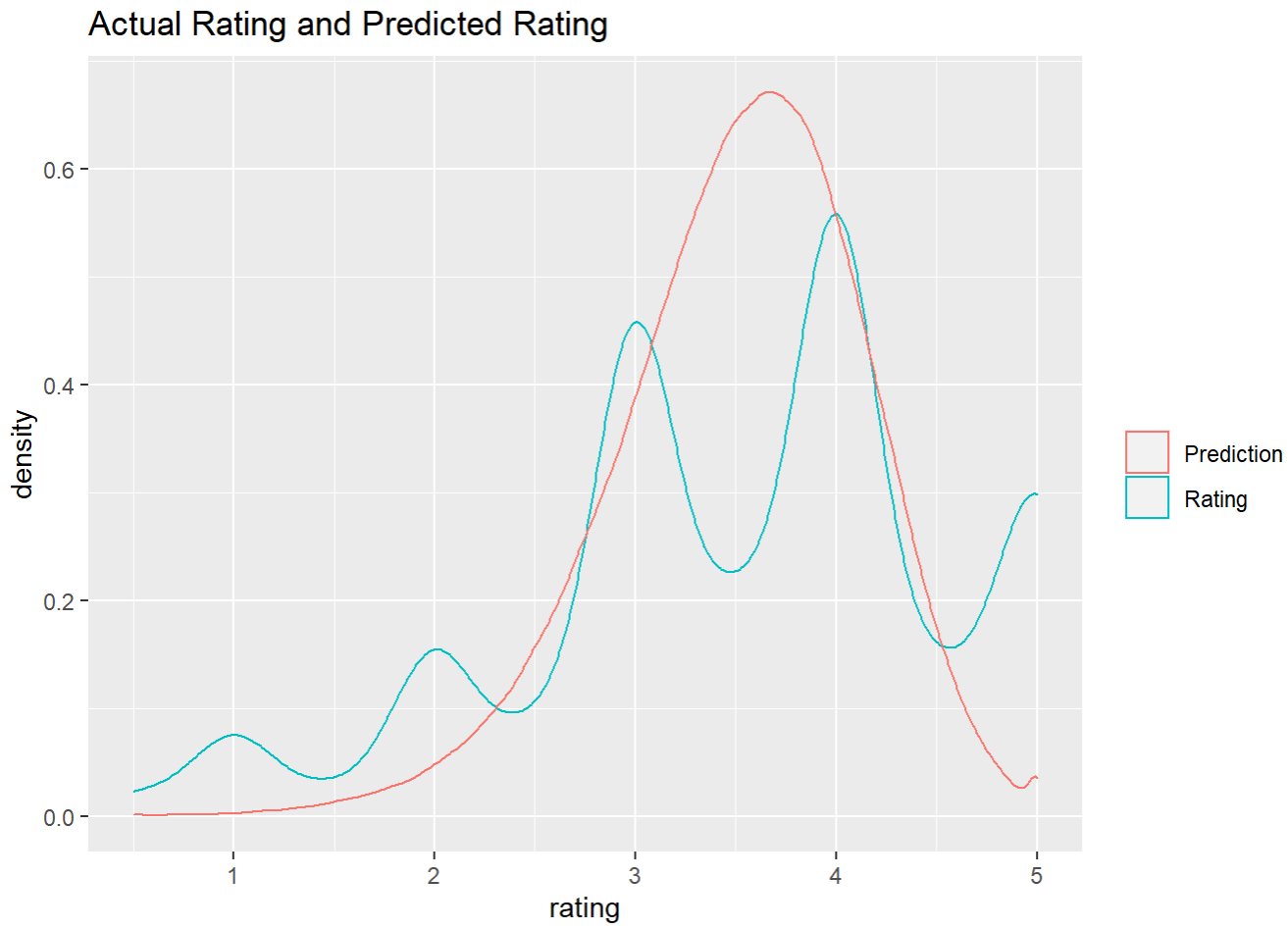

Figure 3: Density Ploy of Predictions and True Ratings

Another potential place for improvement could be in the disentanglement of genres. For example, in the following table it can be seen that the “Mystery” genre is included in many different sets of genres and that these sets have different biases ranging from 2.6 to 4.2. Due to this, it is difficult to say what effect the “Mystery” genre alone has on the rankings. By serperating these genre sets out into something akin to “primary genre”, “secondary genre”, etc. some improvement could be obtained. For example, if “Mystery” was, in fact, a positive bias, it could be possible to weigh the “Mystery” bias more heavily if it occurered sooner in the genre list and less heavily if it were later.

genres	prediction	n
Action|Adventure|Mystery|Sci-Fi	2.607604	902
Action|Adventure|Comedy|Fantasy|Mystery	2.835616	227
Action|Adventure|Fantasy|Mystery	2.893950	229
Action|Crime|Mystery|Thriller	2.973998	496
Crime|Mystery|Romance|Thriller	4.009298	250
Adventure|Mystery|Thriller	4.043622	1672
Crime|Film-Noir|Mystery|Thriller	4.064273	2727
Comedy|Crime|Mystery|Romance|Thriller	4.097039	212
Film-Noir|Mystery|Thriller	4.153011	400

genres	prediction	n
Crime|Mystery|Thriller	4.193597	2993
Film-Noir|Mystery	4.205029	694
Crime|Film-Noir|Mystery	4.222213	429

Finally, the timestamp field, which corresponds to the number of seconds elapsed since January 1, 1970 when the rating was given, was not used. Since the users' locations were not provided in this dataset and since a user's timezone will mean that a given timestamp will correspond to a different time of day for different users, it is not intuitive to see a causal relationship between these timestamps and a rating bias. Hence, this data was ommitted from the analysis. Nonetheless, it is possible that a pattern exists within the timestamp data which could be used to improve the model.