MovieLens Report

Zachary Naumann

4/15/2021

Introduction

The purpose of this report was to determine the effect of various factors on movie ratings with the goal of creating a model that could effectively predict a movie's rating. The relevant dataset, named *edx* and containing about 9 million entries, was contained in a data frame with 6 variables, with each element representing an instance of a movie watched by a user. Below is a sample of the dataset for reference:

```
##    userId movieId rating           timestamp                         title
## 1:      1     122      5 1996-08-02 11:24:06              Boomerang (1992)
## 2:      1     185      5 1996-08-02 10:58:45               Net, The (1995)
## 3:      1     292      5 1996-08-02 10:57:01               Outbreak (1995)
## 4:      1     316      5 1996-08-02 10:56:32               Stargate (1994)
## 5:      1     329      5 1996-08-02 10:56:32 Star Trek: Generations (1994)
## 6:      1     355      5 1996-08-02 11:14:34       Flintstones, The (1994)
##                          genres year month day
## 1:                Comedy|Romance 1996     8   2
## 2:         Action|Crime|Thriller 1996     8   2
## 3:  Action|Drama|Sci-Fi|Thriller 1996     8   2
## 4:       Action|Adventure|Sci-Fi 1996     8   2
## 5: Action|Adventure|Drama|Sci-Fi 1996     8   2
## 6:       Children|Comedy|Fantasy 1996     8   2
```

The purpose of each variable is as follows:
1. *userId* - unique numeric identifier for the user
2. *movieId* - unique numeric identifier for the movie
3. *rating* - rating given by the user to the movie; the lowest rating possible is 0.5, the highest possible is 5
4. *timestamp* - timestamp for when the user watched the movie
5. *title* - title of the movie
6. *genres* - genres assigned to the movie; a movie can have 1 or more genres
7. *year* - year the movie was watched
8. *month* - month the movie was watched
9. *day* - day the movie was watched

There were several key steps required to gather the necessary data. Before doing anything else, the *edx* set was partioning into a training and test set. These were named *train_set* and *test_set* and contained about 7.2 million and 1.8 million entries respectively. The first step after creating and partitioning the dataset was to analyze the individual effects of each variable on the average rating to determine which variables would be viable for the model. After choosing which variables to use, the model was created performing calculations on the biases using a set a training data. The results of these calculations were then used to predict RMSEs with a set of test data, with many lambdas being tested in order to regularize the model. After deciding on the lambda that minimized the RMSE, a final model was calculated using the entire edx

dataset. This model was run against an independent validation set, simply named *validation* and containing about 1 million entries, to come up with the final RMSE value.

Methods/Analysis

Dataset Modifications

Before any data exploration was done, changes were made to the *timestamp* variable to make it more useful for analysis. The original timestamp, which was an unreadable integer, was converted to a human-friendly POSIXct date-time format. The converted timestamp was then split into year, month, and day fields.

Model and Biases

The model used for this recommendation system is represented by the following formula:

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

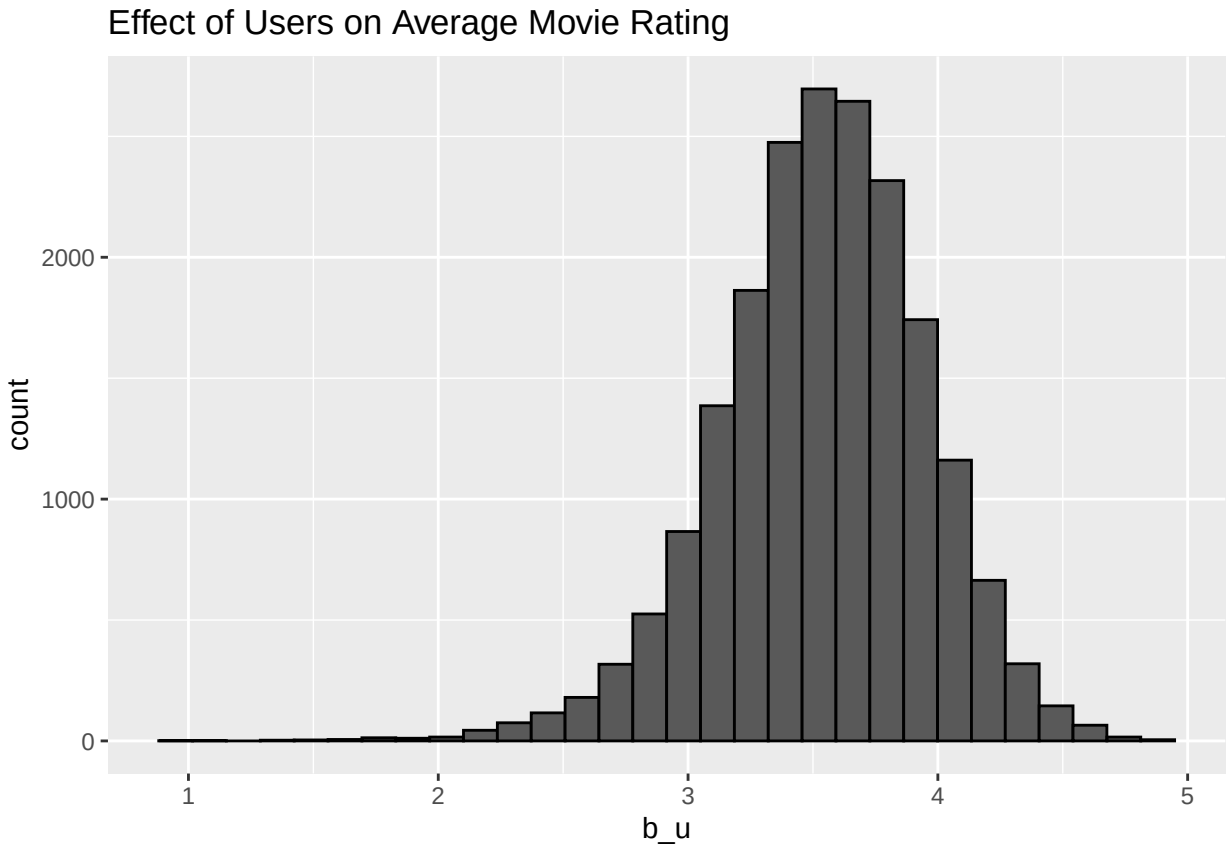
$Y_{u,i}$ is defined as the rating for movie i by user u, μ is the average rating for all movies, $\epsilon_{u,i}$ are the independent errors, and the bs are the biases.

Movie Bias

The movie bias, denoted by the term b_i, represents the effect that the movie itself has on the average rating. This just means that different movies receive different ratings.

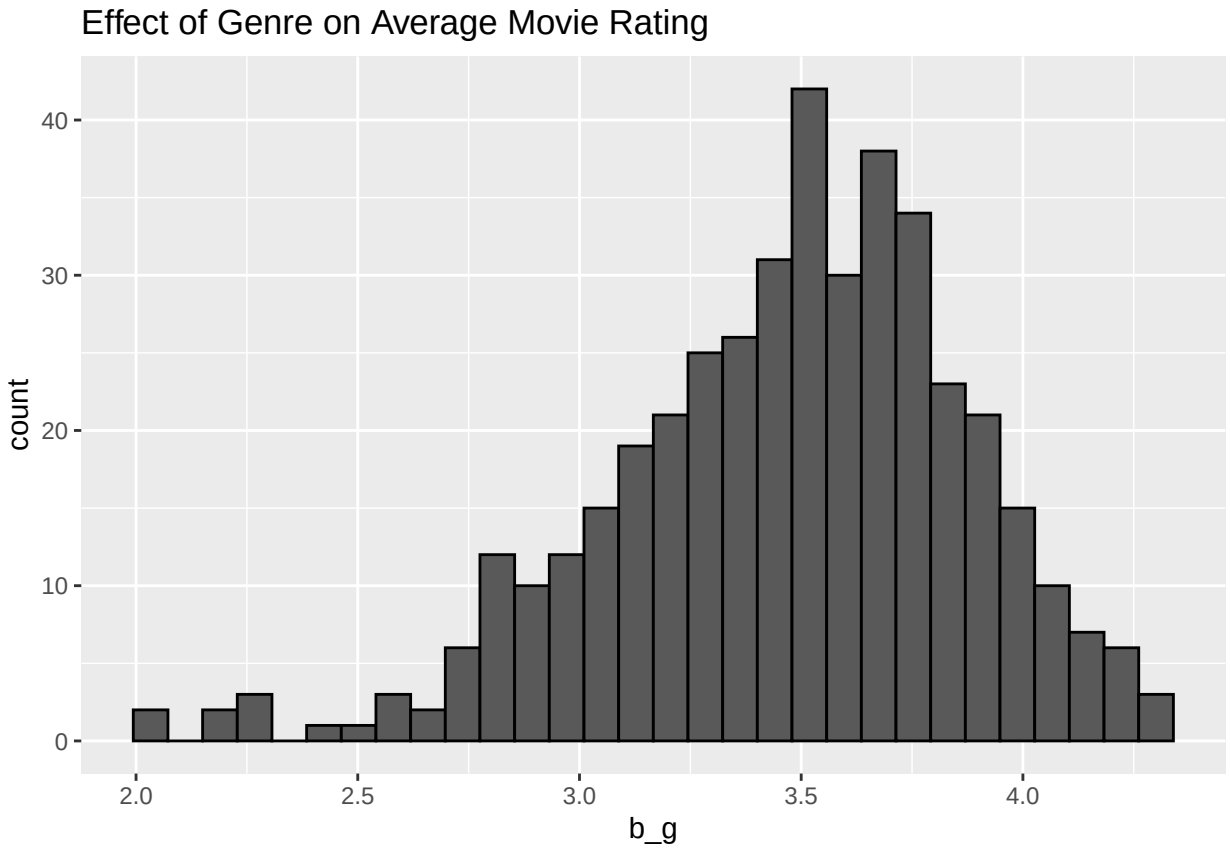
User Bias

The user bias, denoted by the term b_u, represents the effect that a user has on the average rating. Some users tend to be more picky than others, which leads to substantial variation in average movie ratings among users. This can be seen in the graph below:



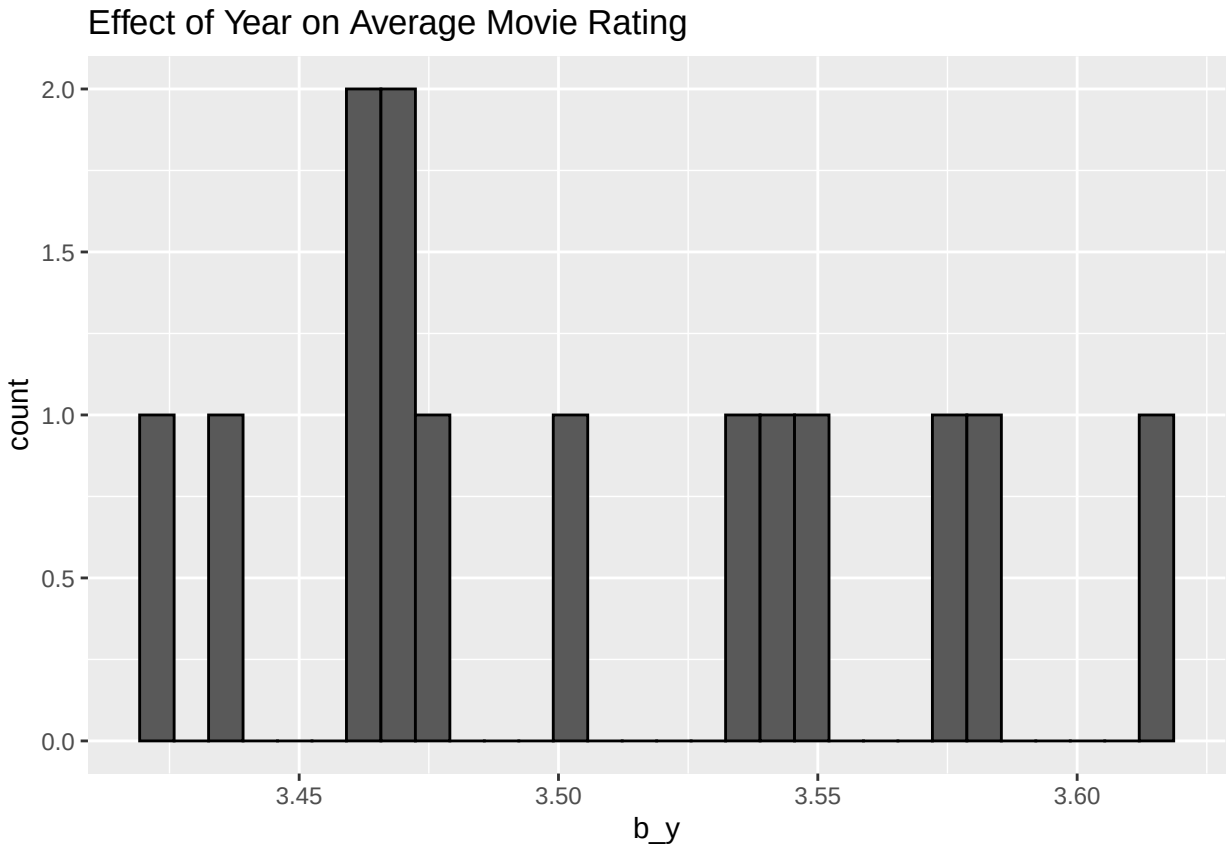
Genre Bias

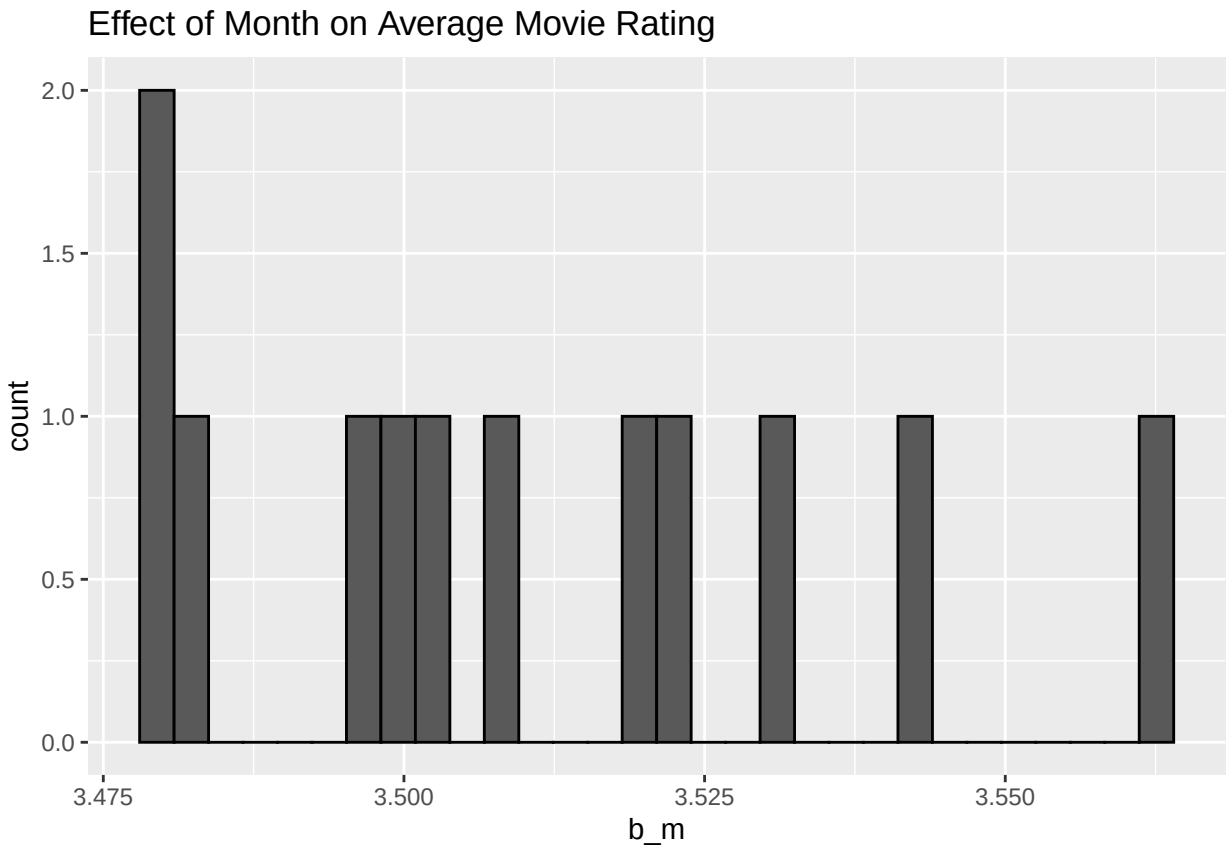
The genre bias, denoted by the term b_g, represents the effect of the movie's genre on the average rating. For this model, a genre can be one or more classifications for a movie (e.g. Action or Comedy|Drama). As with the user bias, there was a significant amount of variability surrounding this field, as shown below:



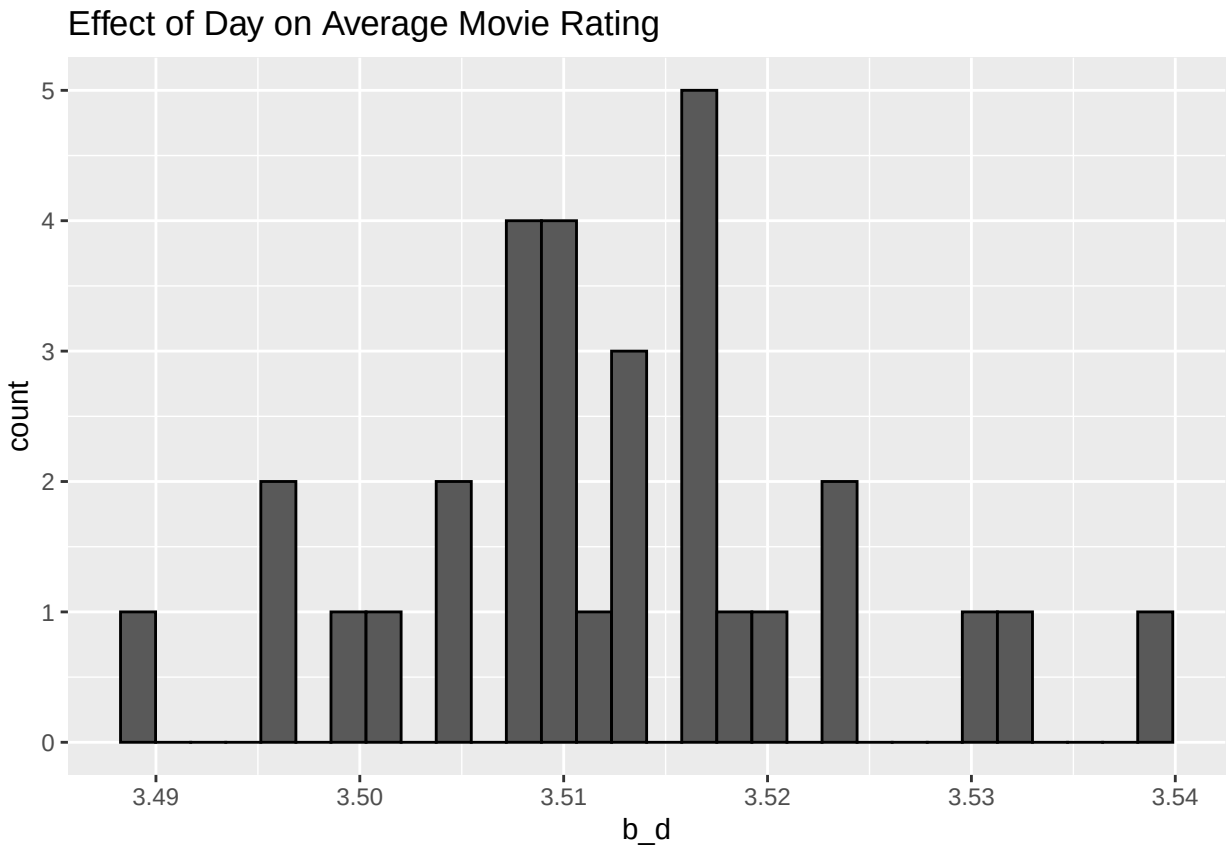
Day Bias

The day bias, denoted by the term b_d, represents the effect that the day of the month has on the average rating. Year, month, and day were all examined to determine variability. For *year* and *month*, there was no clear variability:





There was, however, noticeable variability when analyzing the *day* variable:



As can be seen above, there is a visible normal distribution to the data. It is much narrower than with users or genres, but it was still worth using in the model.

Results

Testing and Regularization

With the model created and the biases chosen, testing was done by performing predictive RMSE calculations on the test set using different lambdas in order to determine which lambda to use in the final regularized model.

The results of the tests produced a minimum RMSE value of 0.8649344 with a corresponding lambda value of 4.75. This lambda value was used in the final calculation. While the minimum RMSE value in these tests was not under the desired threshold of 0.86490, it was close enough to go ahead with the final calculation.

Final RMSE Calculation

The final calculation utilized the entire edx dataset and predicted against the separate validation dataset.

The final RMSE value for the regularized model was 0.8644376. This was well under the desired threshold, which means that this recommendation system performs fairly well at predicting how a movie will be rated.

Conclusion

The results of this report indicate that the model used for this recommendation system was effective at predicting the rating of a movie. There were several fields available in the dataset that were useful as predictors, which greatly helped when building the model. The main limitation when developing the model was the size of the dataset. Even when partitioning the data into training and test sets, the amount of data was into the millions, which reduced the number of modeling methods that were viable. Because of this, any future work on this system would revolve around optimizing how the data is partitioned and referenced in order to improve the speed at which the predictions could be calculated.