MovieLens Project

Bryan Phillips

1/6/2020

Overview

For this project, the task is to create a movie recommendation system using movie ratings data given on a scale from 1-5 from individual users in order to predict user ratings of other movies they have not yet rated.

The data I am using is the 10 million user ratings MovieLens data available at the link below: "https://grouplens.org/datasets/movielens/10m/"

The information given in this dataset includes: User ID, Movie Title, Movie ID, Time Stamp, Genre, Rating

Below is a sample of the data given:

```
##   userId movieId rating timestamp                         title
## 1      1     122      5 838985046              Boomerang (1992)
## 2      1     185      5 838983525               Net, The (1995)
## 4      1     292      5 838983421               Outbreak (1995)
## 5      1     316      5 838983392               Stargate (1994)
## 6      1     329      5 838983392 Star Trek: Generations (1994)
## 7      1     355      5 838984474       Flintstones, The (1994)
##                           genres
## 1                 Comedy|Romance
## 2           Action|Crime|Thriller
## 4  Action|Drama|Sci-Fi|Thriller
## 5        Action|Adventure|Sci-Fi
## 6 Action|Adventure|Drama|Sci-Fi
## 7        Children|Comedy|Fantasy
```

The datasets have already been split into a testing and validation datasets. The testing dataset has 9,000,055 observations from users. The validation dataset has 999,999 observations.

The analysis I plan on using is based off of the recommendation system section of the machine learning course part of the Harvard Data Science Certificate Course. In this section, a matrix factorization model was used to build a recommendation system. I plan on building off of this original analysis.

The method in which I am testing the validity of my recommendation is by seeing the Root Mean Squared Error (RMSE) of each model on the validation dataset provided. My goal is to get a RMSE of less than 0.8649.

note: The datasets used in this analysis were provided on the group forum DropBox folder because the datasets generated from the given dataset code did not match the grader results. For consistency, I used the given datasets throughout the analysis.

Below is a link to the dropbox folder of the datasets I used and reference in my code https://www.dropbox.com/sh/n2d1gji7kkdsvhi/AADnvvXmRqXOfoP7bHN_1ydda?dl=0

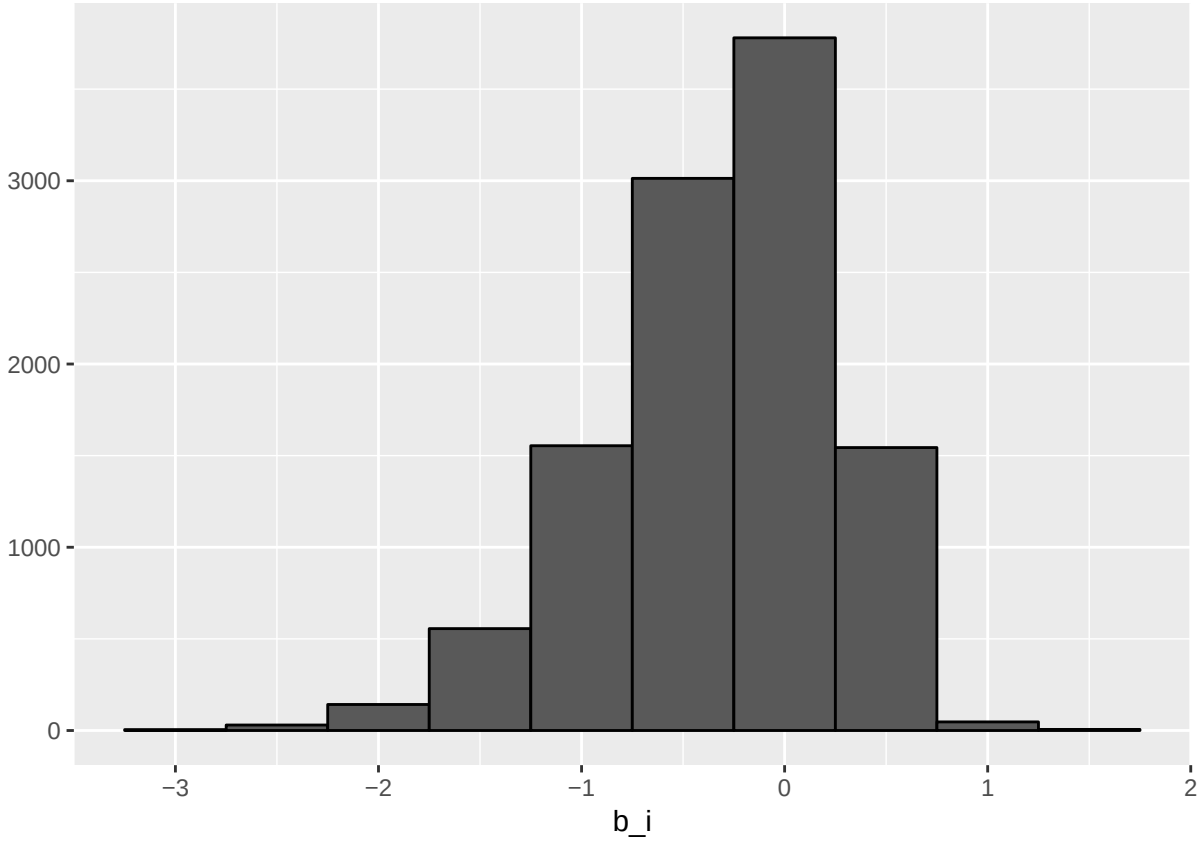
All the code used for this analysis is available on my GitHub repository linked below: https://github.com/bkphillips/Movie_rating_prediction/

Analysis

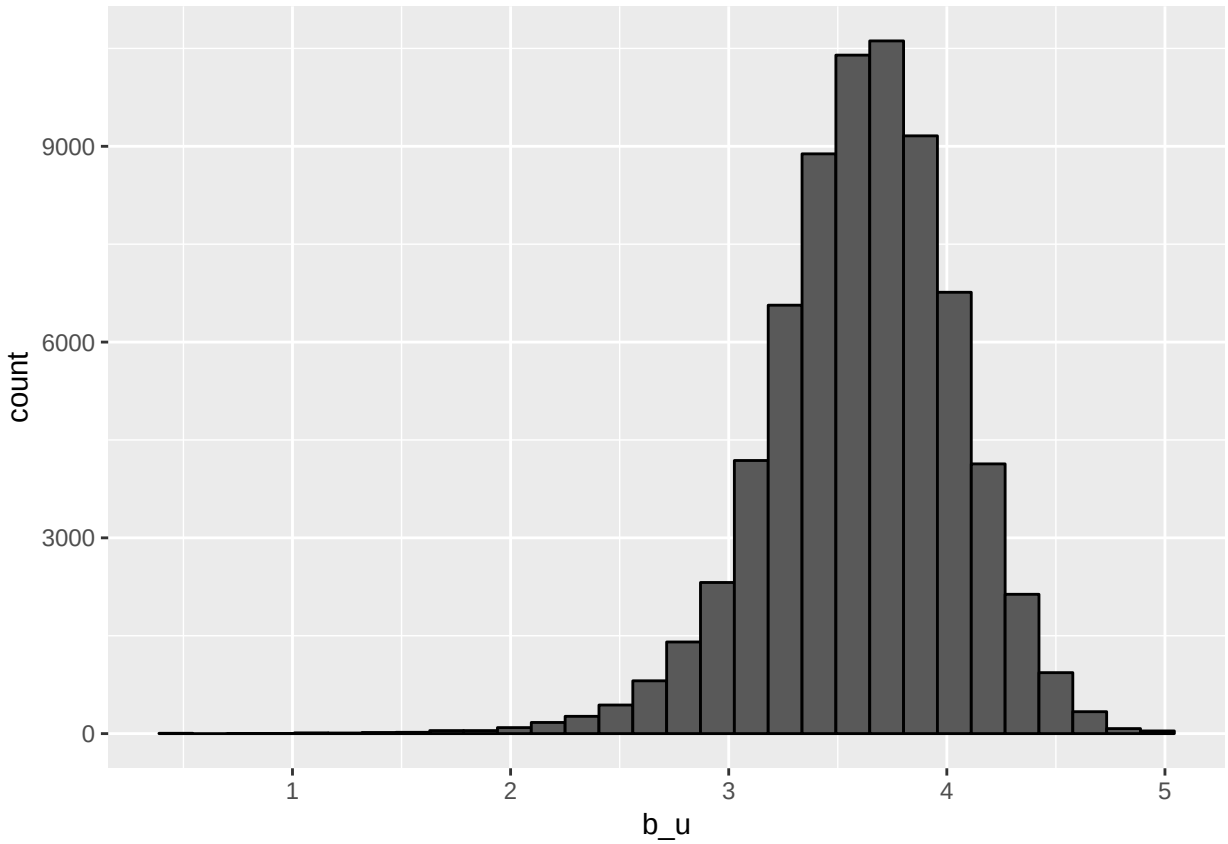
For this analysis, I build off the original recommendation system analysis previously mentioned in the machine learning course.

The first model looks just at the overall average of the test dataset, which is 3.512465. This naive model has a RMSE of 1.0603.

The second model is based off the observation that each individual movie has an effect on the rating. This movie effect is represented by "b_i" in the second movel. Below is a distribution of this effect:

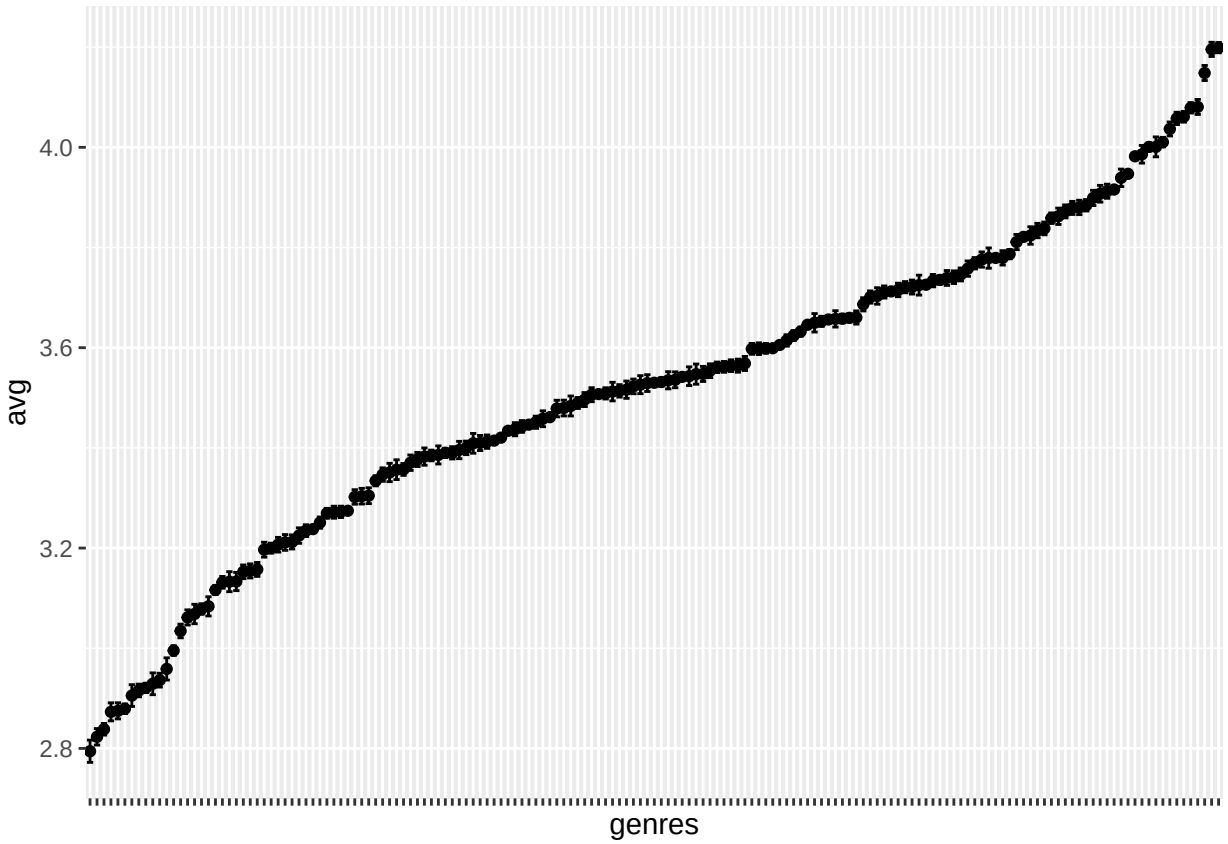


The third model is based off of the observation that there is a user bias that affects the ratings. Some users may be very picky, while others are very generous in their ratings of movies. Below is a distribution of user ratings:

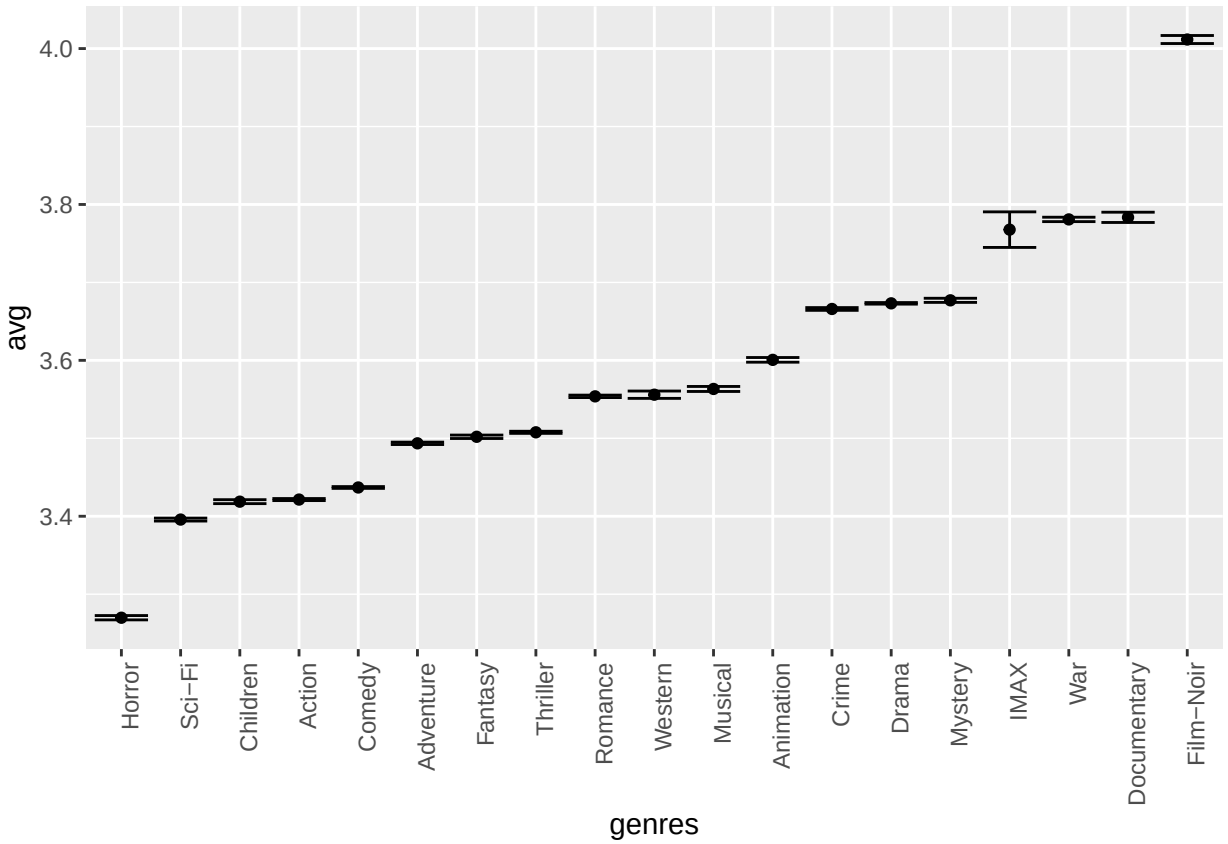


The third model combines both the movie and user effects as b_i + b_u.

In my own analysis, I wanted to incorporate the genre into the model in order to get a more accurate recommendation. In order to do this, I converted the genre information into a tidy format in order to provide a more accurate recommendation. Before converting the genre, there were 797 distinct categories for genre shown in the average rating distribution below:



I then converted the data into a long format where I now have only 20 different genres that can more accurately predict the rating, as shown by the plot below of the average rating by the new tidy genres:



For the fourth model, I used this genre effect (b_g) seen in the figure above alongise both the movie and user effect represented as "b_i + b_u + b_g".

For a fifth model, I made an assumption that each user most likely has a strong preference for a genre that will affect the rating, which would help make a strong prediction. I represent is user genre effect as b_ug in a model that incorporates also the movie and user effects as "b_i + b_u + b_ug".

When validating this last model on the validation data, I noticed that there is not always User Genre rating information for each user in the new dataset, which in that case the model assumes just the movie and user effects.

Results

Below are my RMSE results from the test data on each model I previously described above:

method	RMSE
Just the average	1.0603313
Movie Effect Model	0.9423475
Movie + User Effects Model	0.8567039
Movie + User Effects + Genre Effects Model	0.8565891
Movie + User Effects + User Genre Effects Model	0.8097699

It appears that the last model that incorporated the user-genre effects had an significant improvement on the RMSE on the testing data.

For validation, I only looked at the last three models on the validation dataset. Below are the results of validating these three models using the validation dataset:

method	RMSE
Movie + User Effects Model	0.8653488
Movie + User Effects + Genre Effects Model	0.8652323
Movie + User Effects + User Genre Effects Model	0.8497552

Conclusion

It appears that the fifth model of Movie, User, and User Genre Effects (b_i + b_u + b_ug) had the strongest predictive performance on the validation dataset with a RMSE of 0.8497552. It is interesting that addition of the genre effects on the model did not cause much predictive improvement on the model.

The results of this analysis show that individual tastes for a particual genre have a significant impact on their rating for a particular movie. In the real world, this makes sense given that many people have a strong affinity towards a particular genre of movies.

A major limitation of this method is that it requires previous user information and would not have much predictive performance on users that had not already provided rating information into the model.