MovieLens Project

Myint Moe Chit

13/06/2019

Introduction

One of the applications of machine learning is predicting the response of a consumer as accurate as possible. In this project, I attempt to develop a model that predict the rating given by user on a movie using the methodology described in Irizarry (2019). To develop the basic machine learning model, I use a subset of the database generated by GroupLens research lab. The dataset includes 10 million ratings for 10677 movies by 69878 users.

The goal of the project is to develop a machine learning algorithm that can predict movie rating relatively accurately. The accuracy of the algorithm is measured by the residual mean squared error (RMSE) on a test (validation) dataset. The target accuracy of the project is to achieve an RMSE score lower than 0.8775. Since RMSE can be interpret as a typical predicting error, the target is to achieve the margin of error lower than 0.87 star.

To achieve the targeted value of RMSE, I train a machine learning algorithm using the inputs in a subset of the data (a train set with about 9 million observations) to predict movie ratings in the validation set (about 1 million observations).

Data preparation and method of analysis

The dataset used in this project is download from http://files.grouplens.org/datasets/movielens/ml-10m.zip and generate a data frame called "movielens" in R. The summary of data in tidy format is presented below.

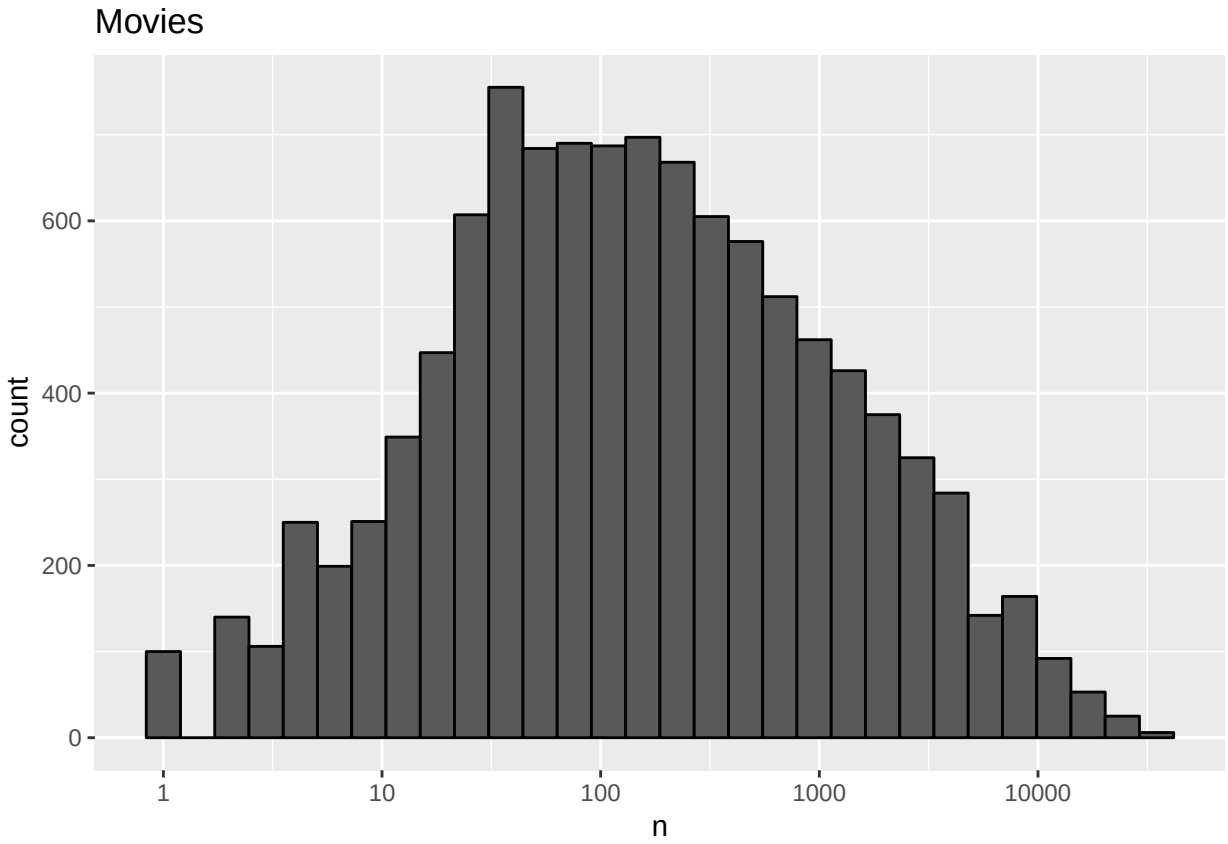
```
## # A tibble: 10,000,054 x 6
##    userId movieId rating timestamp title            genres
##     <int>   <dbl>  <dbl>     <int> <chr>            <chr>
##  1      1     122      5 838985046 Boomerang (1992) Comedy|Romance
##  2      1     185      5 838983525 Net, The (1995)  Action|Crime|Thriller
##  3      1     231      5 838983392 Dumb & Dumber (~ Comedy
##  4      1     292      5 838983421 Outbreak (1995)  Action|Drama|Sci-Fi|T~
##  5      1     316      5 838983392 Stargate (1994)  Action|Adventure|Sci-~
##  6      1     329      5 838983392 Star Trek: Gene~ Action|Adventure|Dram~
##  7      1     355      5 838984474 Flintstones, Th~ Children|Comedy|Fanta~
##  8      1     356      5 838983653 Forrest Gump (1~ Comedy|Drama|Romance|~
##  9      1     362      5 838984885 Jungle Book, Th~ Adventure|Children|Ro~
## 10      1     364      5 838983707 Lion King, The ~ Adventure|Animation|C~
## # ... with 10,000,044 more rows
```

The data frame includes five variables: • userId: ID for unique user • movieId: ID for a movie • rating: Rating given to a movie (minimum 0.5 star to maximum 5 star) • time stamp: Time and data in which the rating was provided • title: Title of the movie • genres: Type of movie

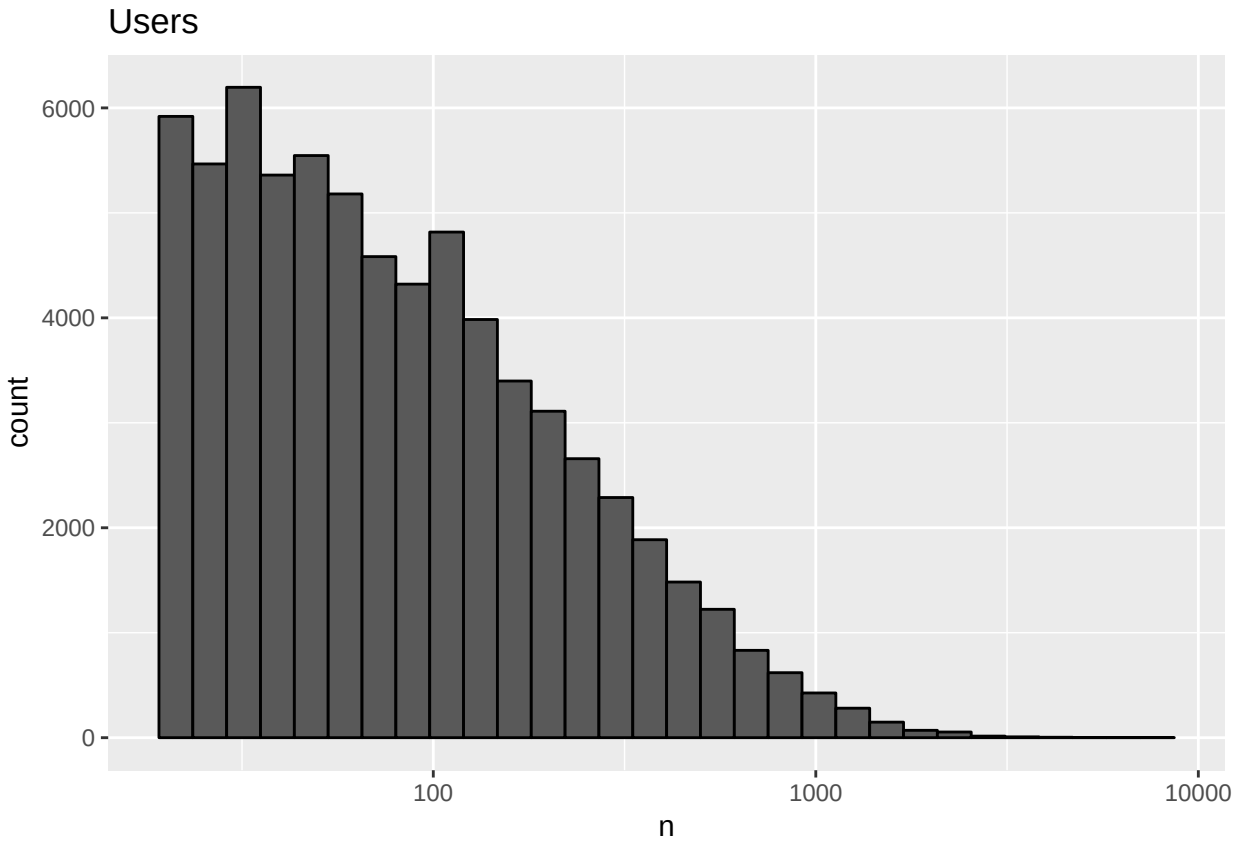
The data set includes 69878 users who rated 10677 movies.

To identify the factors that influence a user's rating on a movie, first we analyse the distribution of the data.

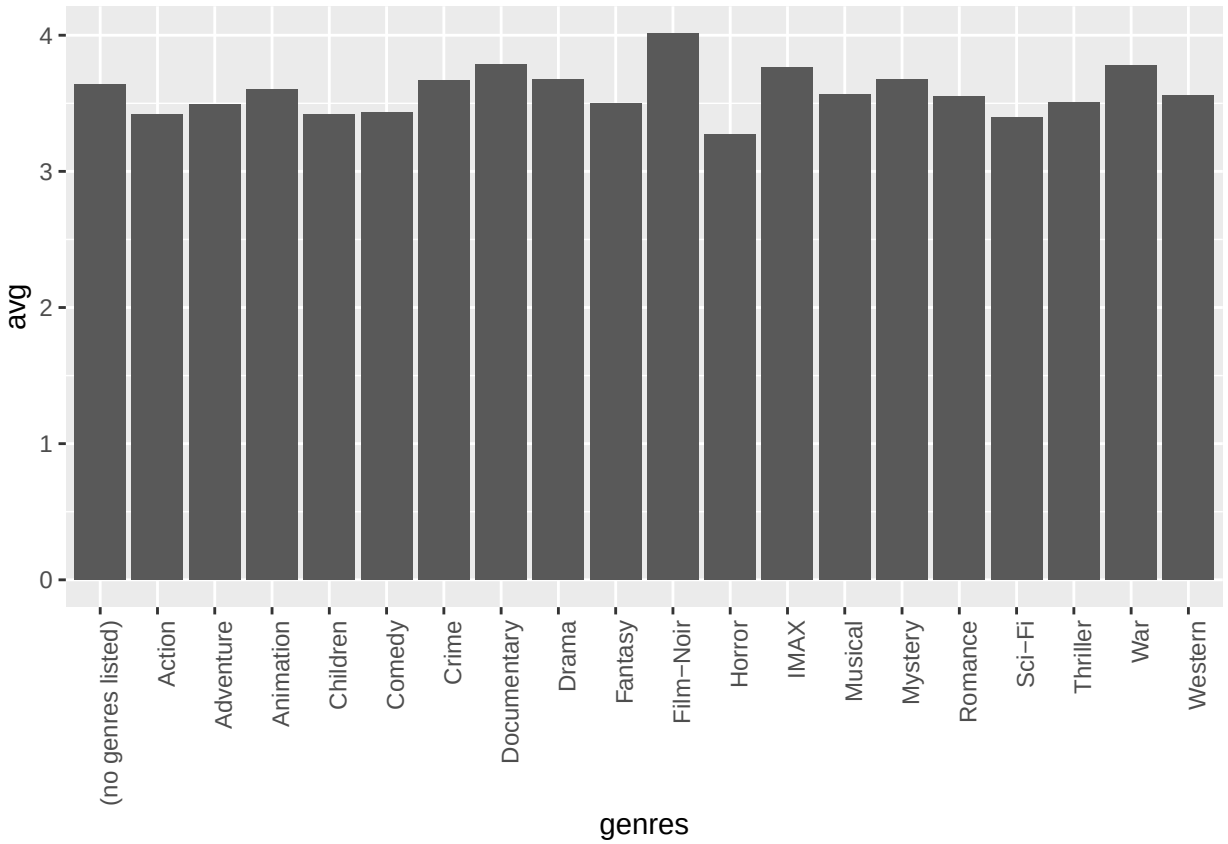
The following distribution indicates that the different movies the number of ratings is different across the movies, indicating rating is associated with the movie.



Then, we plot the distribution of rating across unique users. The following distribution suggests the number of rating given are also different across unique users.



The data set also includes a variable that indicates the genre of the movie and some movies fall under several genres. To evaluate the association between the genre and a rating given on a movie, we first separate the genres and assign a movie to each genre. Then, we plot the average rating given to movies fall under different genre. Again, the following plot show the average rating differs across different genres suggesting rating is also influenced by the genre of a movie.



The plots presented above clearly show strong evidence of user, movie, genre effect.

To predict movie rating, I divide the data into training set and validation (test) set. The training set (90 percent of the observations) is used to build the algorithm which is then applied on the validation set (10% of the observations) to access the accuracy.

The summary of training (edx) set and validation set in tidy format are presented below.

```
## # A tibble: 9,000,055 x 6
##    userId movieId rating timestamp title              genres
##  *  <int>   <dbl>  <dbl>     <int> <chr>              <chr>
##  1      1     122      5 838985046 Boomerang (1992)   Comedy|Romance
##  2      1     185      5 838983525 Net, The (1995)    Action|Crime|Thril~
##  3      1     292      5 838983421 Outbreak (1995)    Action|Drama|Sci-F~
##  4      1     316      5 838983392 Stargate (1994)    Action|Adventure|S~
##  5      1     329      5 838983392 Star Trek: Generat~ Action|Adventure|D~
##  6      1     355      5 838984474 Flintstones, The (~ Children|Comedy|Fa~
##  7      1     356      5 838983653 Forrest Gump (1994) Comedy|Drama|Roman~
##  8      1     362      5 838984885 Jungle Book, The (~ Adventure|Children~
##  9      1     364      5 838983707 Lion King, The (19~ Adventure|Animatio~
## 10      1     370      5 838984596 Naked Gun 33 1/3: ~ Action|Comedy
## # ... with 9,000,045 more rows
```

```
## # A tibble: 999,999 x 6
##    userId movieId rating  timestamp title                genres
##     <int>   <dbl>  <dbl>      <int> <chr>                <chr>
##  1      1     231      5  838983392 Dumb & Dumber (1994)  Comedy
##  2      1     480      5  838983653 Jurassic Park (1993)  Action|Adventur~
##  3      1     586      5  838984068 Home Alone (1990)     Children|Comedy
```

```
##  4      2     151    3    868246450 Rob Roy (1995)        Action|Drama|Ro~
##  5      2     858    2    868245645 Godfather, The (1972) Crime|Drama
##  6      2    1544    3    868245920 Lost World: Jurassic~ Action|Adventur~
##  7      3     590    3.5 1136075494 Dances with Wolves (~ Adventure|Drama~
##  8      3    4995    4.5 1133571200 Beautiful Mind, A (2~ Drama|Mystery|R~
##  9      4      34    5    844416936 Babe (1995)           Children|Comedy~
## 10      4     432    3    844417070 City Slickers II: Th~ Adventure|Comed~
## # ... with 999,989 more rows
```

The accuracy of the model developed to predict movie rating will be measured by the residual mean squared error (RMSE) on the validation set.

The function that computes the RMSE for vectors of ratings and their corresponding predictors is described below:

```
RMSE <- function(true_ratings, predicted_ratings){
    sqrt(mean((true_ratings - predicted_ratings)^2))
}
```

The RMSE obtained from prediction can be interpreted as the typical error made in predicting a movie rating. If this number is larger than 1, the typical error made in prediction is larger than one star. The aim of this project is to build an algorithm that yields an RMSE lower than 0.8775.

Analysis and Results

Following the simplified version of Normalisation of Global Effect method used in Irizarry (2019), I decompose the rating given by a user on a movie into four different components: the baseline rating (the average of all user-movie ratings), user-specific effect, movie-specific effect, and genre-specific effect.

First, I calculate the average rating of the users in the sample. The average rating is 3.51 star.

If we assume that the expected rating on a movie is 3.51 and the differences (lower of higher than the expected rating) were due to random variation, the accuracy of this prediction can be measured by RMSE. As shown in the following table, RMSE of the model that use the average value to predict the movie rating is 1.06, suggesting the margin of estimation error is about 1 star.

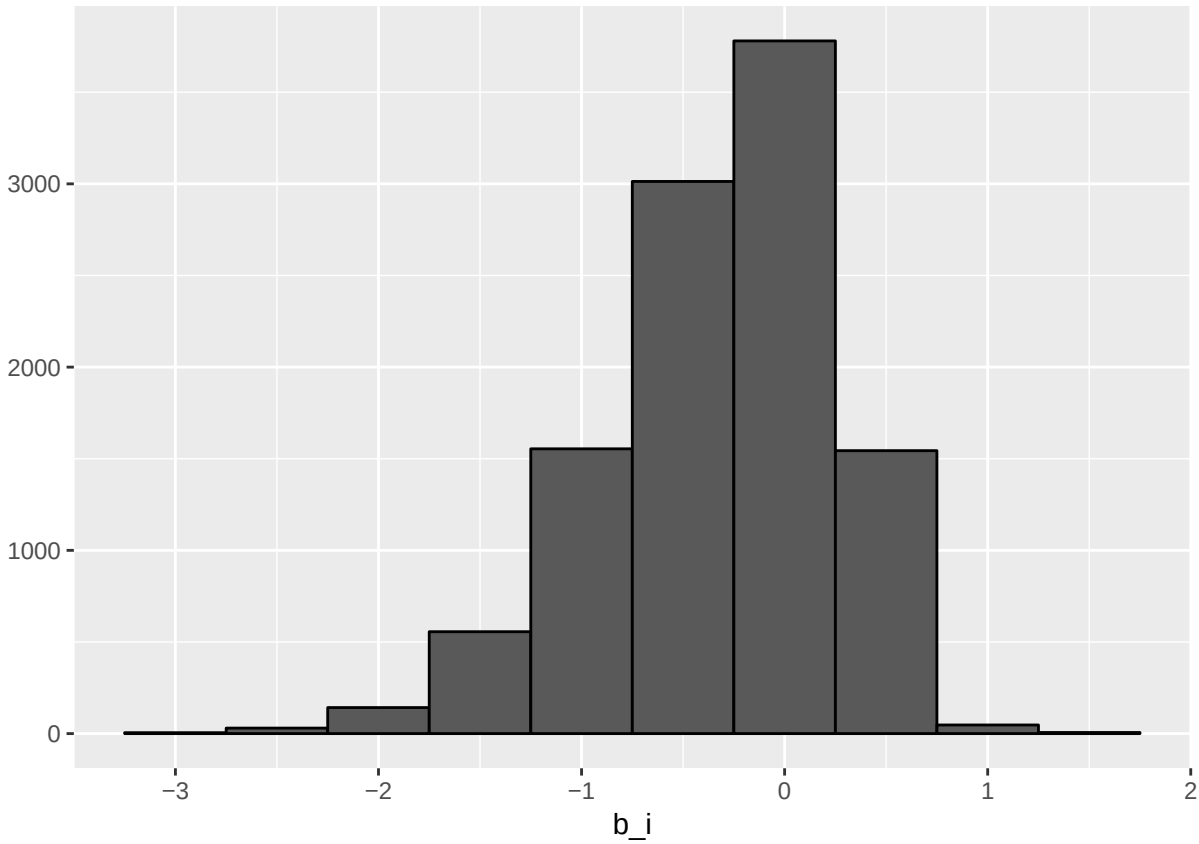
```
## [1] 1.061202
```

Method	RMSE
Just using the average	1.061202

Model with movie-specific bias

As demonstrated in the previous section, different movies get different rating and some movies get rated more than others. To develop a model for more accurate estimation, I calculate the movie-specific bias as outlined in Irizarry (2019). The movie-specific bias is calculated as the average of the differences between individual user's rating on a specific movie and the expected (average) rating of all users-movies. The estimated movie-specific bias is calculate using the training data set.

As shown in the following plot, movie-specific bias vary significantly with a minimum of – 3 stars to a maximum of 1.5 stars.

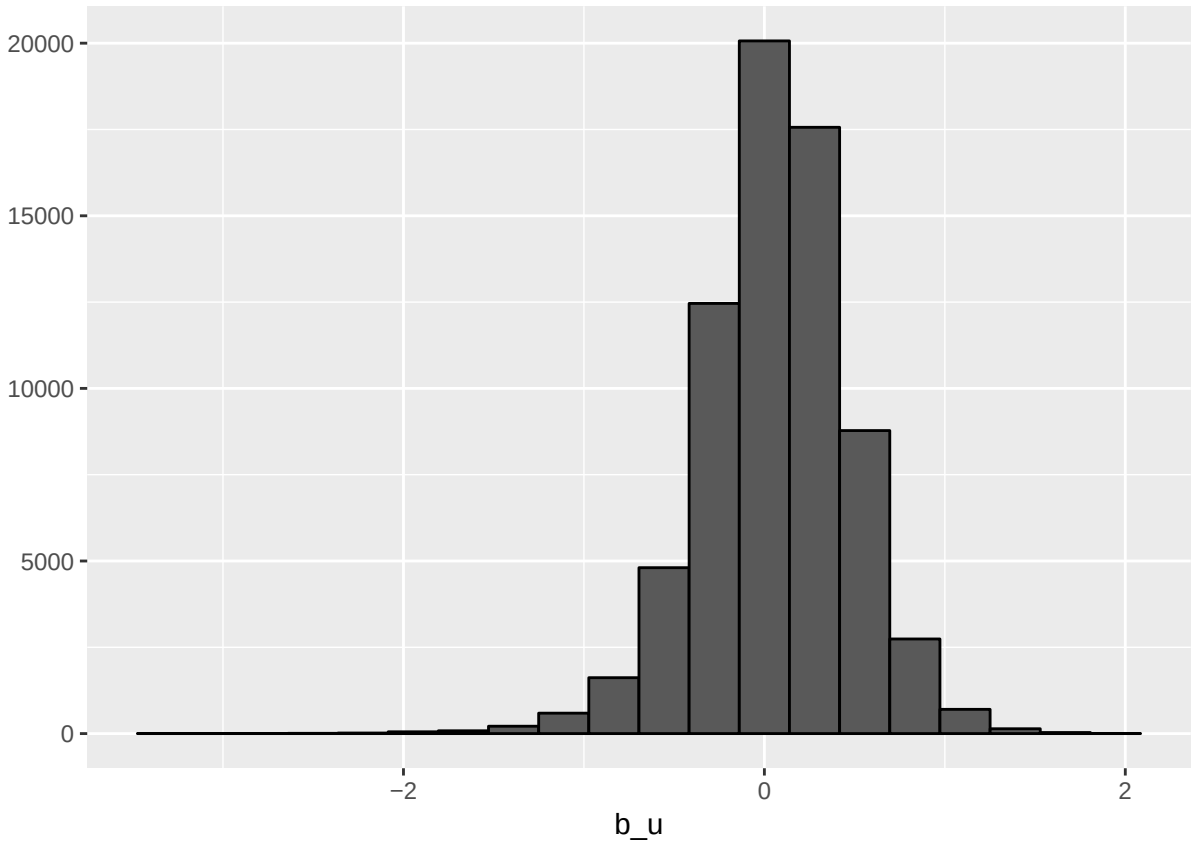


I construct a model that considers movie-specific bias in predicting movie rating and test its accuracy on the validation data set. The obtained RMSE indicates that there is an improvement in the prediction. RMSE decreased from 1.06 to 0.9439.

Method	RMSE
Just using the average	1.0612018
Movie-specific Effect Model	0.9439087

Model with Movie and User-specific bias

The user-specific bias is calculated as the average of the differences between individual user's rating on a specific movie and the sum of expected (average) rating of all users-movies plus movie-specific bias. The estimated user-specific bias is calculate using the training data set. As shown in the following plot, movie-specific bias vary significantly with a minimum of – 2 stars to a maximum of 1.8 stars.



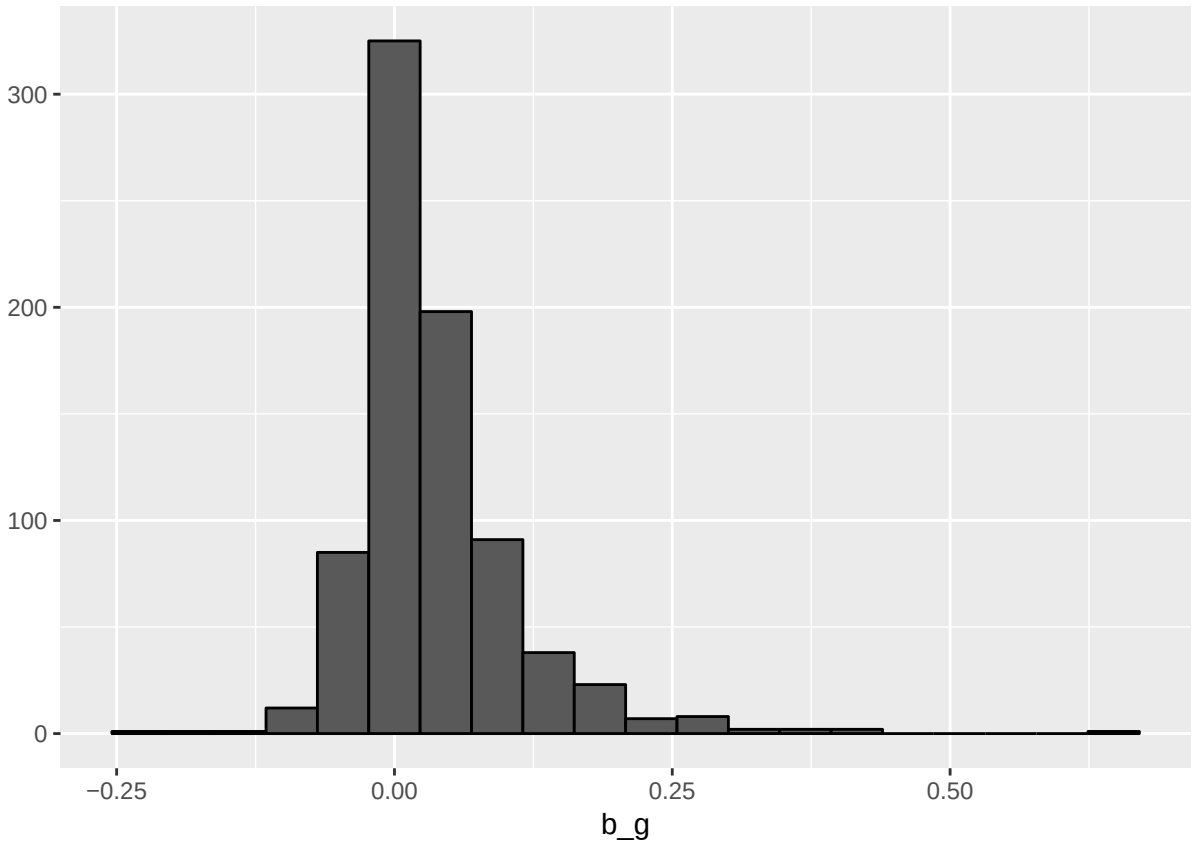
Then, I construct a prediction model that take account of the movie-specific and user-specific bias using the validation data. As shown in the table, RMSE of this approach (0.8653) has been improved and lower than that of the model with only movie-specific bias.

Method	RMSE
Just using the average	1.0612018
Movie-specific Effect Model	0.9439087
Movie + User-specific Effects Model	0.8653488

Model with Movie, User, and Genre-specific bias

Since the genre of a movie also influenced the movie rating, I develop a model with all three biases. The genre-specific bias is calculated as the average of the differences between individual user's rating on a specific movie and the sum of expected (average) rating of all users-movies plus movie and user-specific biases. The estimated genre-specific bias is calculate using the training data set.

As shown in the following plot, movie-specific bias vary with a minimum of – 0.2 stars to a maximum of 0.6 stars.



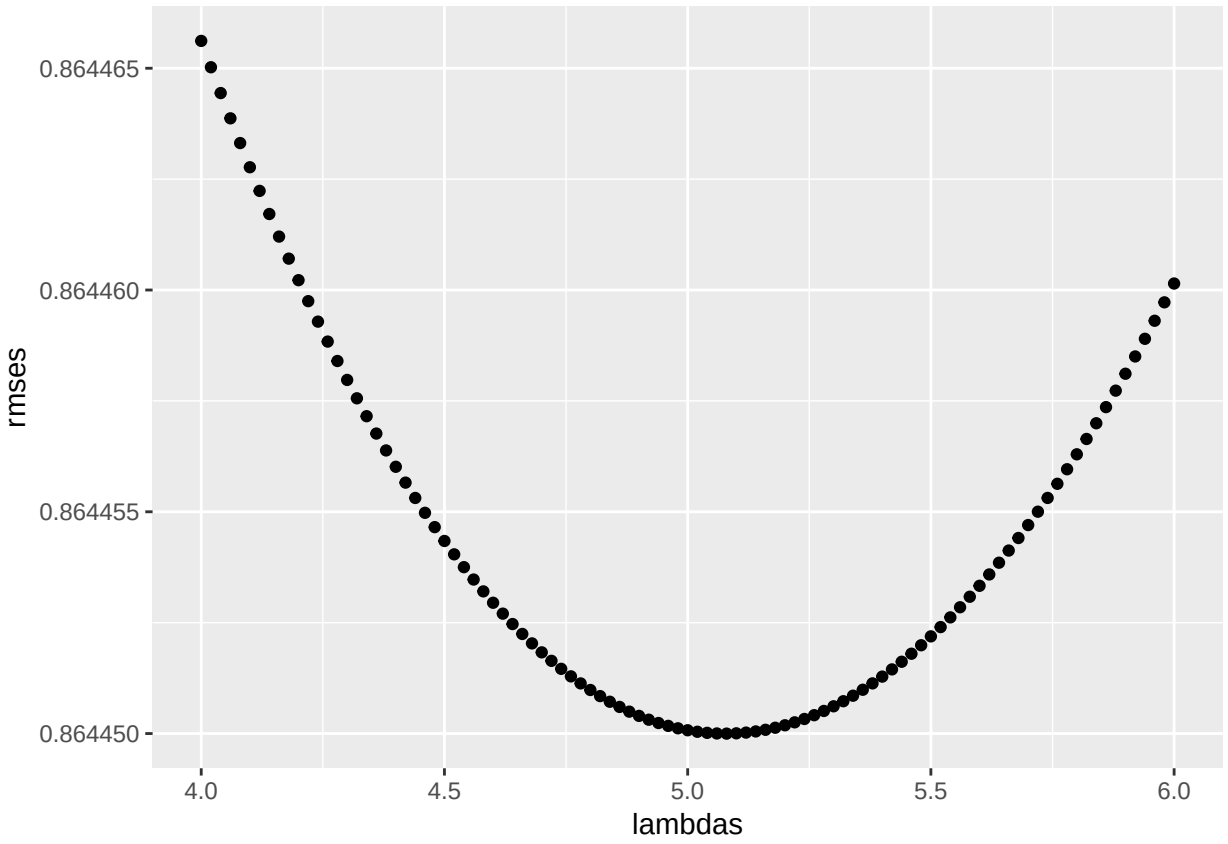
Then, I construct predictors using the validation data. As shown in the table, RMSE decreases to 0.8649.

Method	RMSE
Just using the average	1.0612018
Movie-specific Effect Model	0.9439087
Movie + User-specific Effects Model	0.8653488
Movie + User + genres-specific Effects Model	0.8649469

Regularization

As suggested in Irizarry (2019). there could be potential uncertainty in our prediction due to the influence of the highest and the lowest ratings given to movies watched by very few users. To control for those noisy estimates that could affect our prediction, I apply regularisation method that penalise large estimates that are formed using small sample sizes. The penalty parameter, Lambda, is selected using a cross-validation method.

As shown in the above plot the value of Lambda that minimise RMSE is 5.08.



Then, I calculate the RMSE of the regularised model that control movie, user, and genres-specific effects. As shown in the following result table, the penalized estimates provide an improvement over the previous estimates.

Method	RMSE
Just using the average	1.0612018
Movie-specific Effect Model	0.9439087
Movie + User-specific Effects Model	0.8653488
Movie + User + genres-specific Effects Model	0.8649469
Regularized Movie + User + Genres Effect Model	0.8644500

Conclusion

I develop a machine learning model to predict the rating on a movie given by users using a dataset generated by GroupLens research lab. I use a simplified version of Normalisation of Global Effect method. After controlling the user-specific, movie-specific, and genre specific effects and outlier ratings given to a movie watched by very few people, the accuracy of the model measured by RMSE is 0.864. There are a number of limitations in this project. First, the model is based on a simplified version of machine learning model. Second, the interaction between users and movies (i.e., certain movies might be rated differently by some group of users) should be included to further improve the accuracy.

Reference

Irizarry, R. A. (2019) Introduction to Data Science: Data Analysis and Prediction Algorithms with R (available at: https://rafalab.github.io/dsbook/)