Movie Recommendation System

Sadia Boksh

29/01/2021

Introduction

For this project, I will be creating a movie recommendation system using the 10M verson of the MovieLens data set. Here I will train a machine learning model using the inputs in one subset to predict movie ratings in the validation set.

Multiple models to be fit for this purpose and their RMSEs will be calculated. I will start with a simple model that uses the movie average for prediction. Eventually, I will modify the model to include movie effects, user effects and genre effects. In addition, I will use regularization to add penalty terms to shrink the effect of smaller sample sizes towards 0. The model that produces the least RMSE will be used on the validation data set to calculate the final RMSE.

Analysis

First I will do some basic data exploration. The MovieLens data set will be initially split into 10% validation set for calculating final RMSE and 90% training set to train and test my models and will be called edx set.

Number of rows and cols in the edx set:

```
## [1] 9000055
```

```
## [1] 6
```

Number of rows and cols in the validation set:

```
## [1] 999999
```

```
## [1] 6
```

Number of zeros as ratings:

```
##    n
## 1: 0
```

Number of 3's as rating:

```
##          n
## 1: 2121240
```

Number of different movies:

```
## [1] 10677
```

Number of different user:

```
## [1] 69878
```

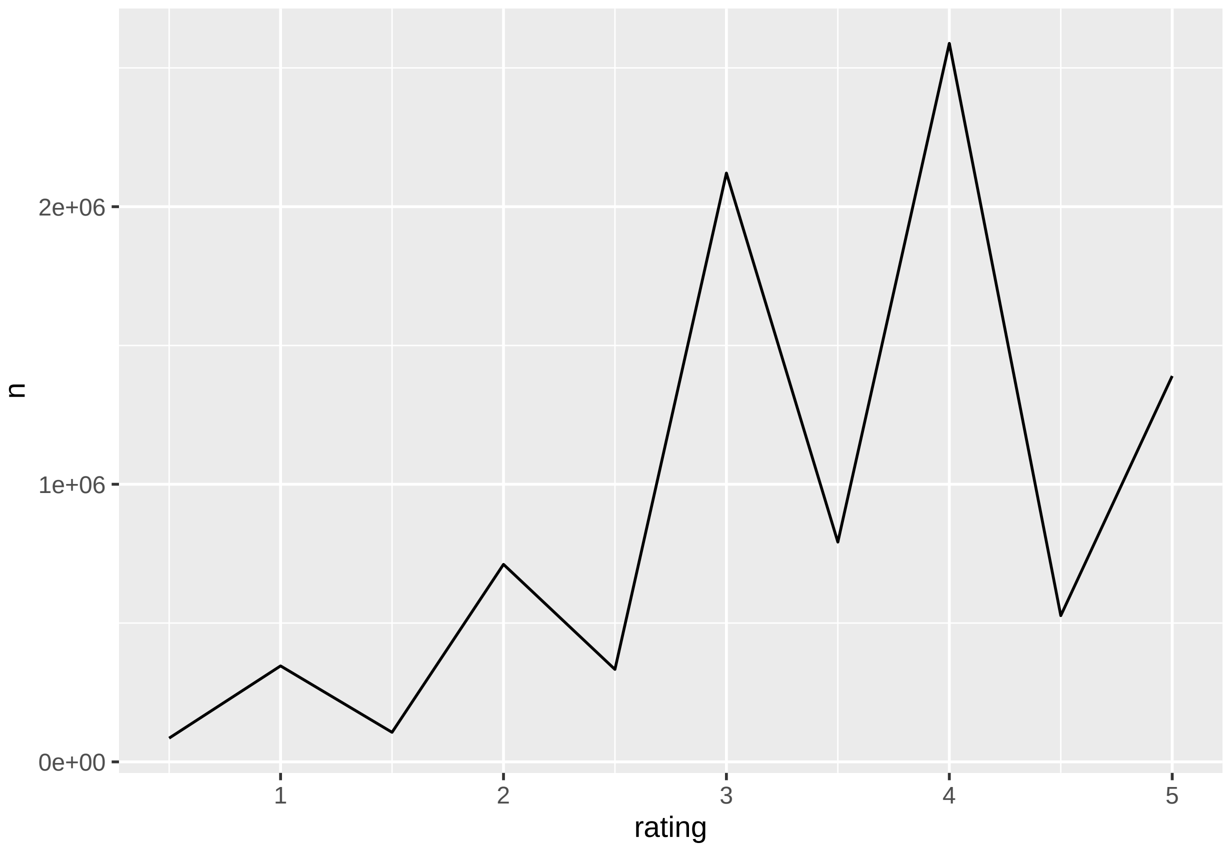
Number movie ratings in Drama, Comedy, Thriller, Romance in edx dataset:

```
##      Genre       n
## 1    Drama 3910127
## 2   Comedy 3540930
## 3 Thriller 2325899
## 4  Romance 1712100
```

Highest rated movies:

```
## # A tibble: 10,677 x 3
##    movieId     n title
##      <dbl> <int> <chr>
##  1     296 31362 Pulp Fiction (1994)
##  2     356 31079 Forrest Gump (1994)
##  3     593 30382 Silence of the Lambs, The (1991)
##  4     480 29360 Jurassic Park (1993)
##  5     318 28015 Shawshank Redemption, The (1994)
##  6     110 26212 Braveheart (1995)
##  7     457 25998 Fugitive, The (1993)
##  8     589 25984 Terminator 2: Judgment Day (1991)
##  9     260 25672 Star Wars: Episode IV - A New Hope (a.k.a. Star Wars) (1977)
## 10     150 24284 Apollo 13 (1995)
## # ... with 10,667 more rows
```

Greatest number of ratings:



From the plot above we can say that the half star ratings are less common than full star ratings.

Model Fitting

In order to train and test my model, I will further split up the edx dataset into 20% test set and 80% train set.

Number of rows in test and train set respectively:

```
## [1] 1799968
```

```
## [1] 7200087
```

First Model

For the first model, we will use just the average of all movies for prediction. RMSE for this simple model in the test set:

```
## [1] 1.060561
```

Movie Effect

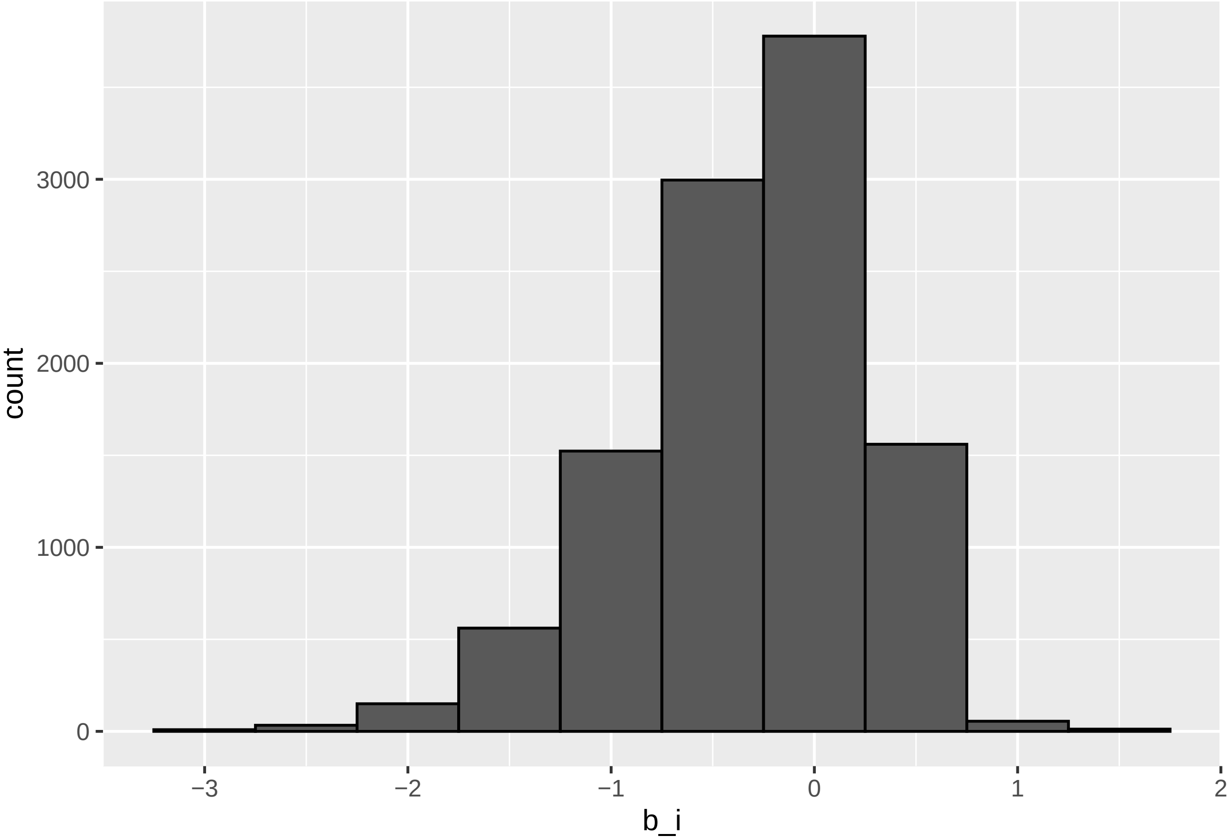
Next I will cater for the movie effect. For movie effect my model is:

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

and RMSE for this model:

```
## [1] 0.9439868
```

We can also visualize the movie effect:



User Effect

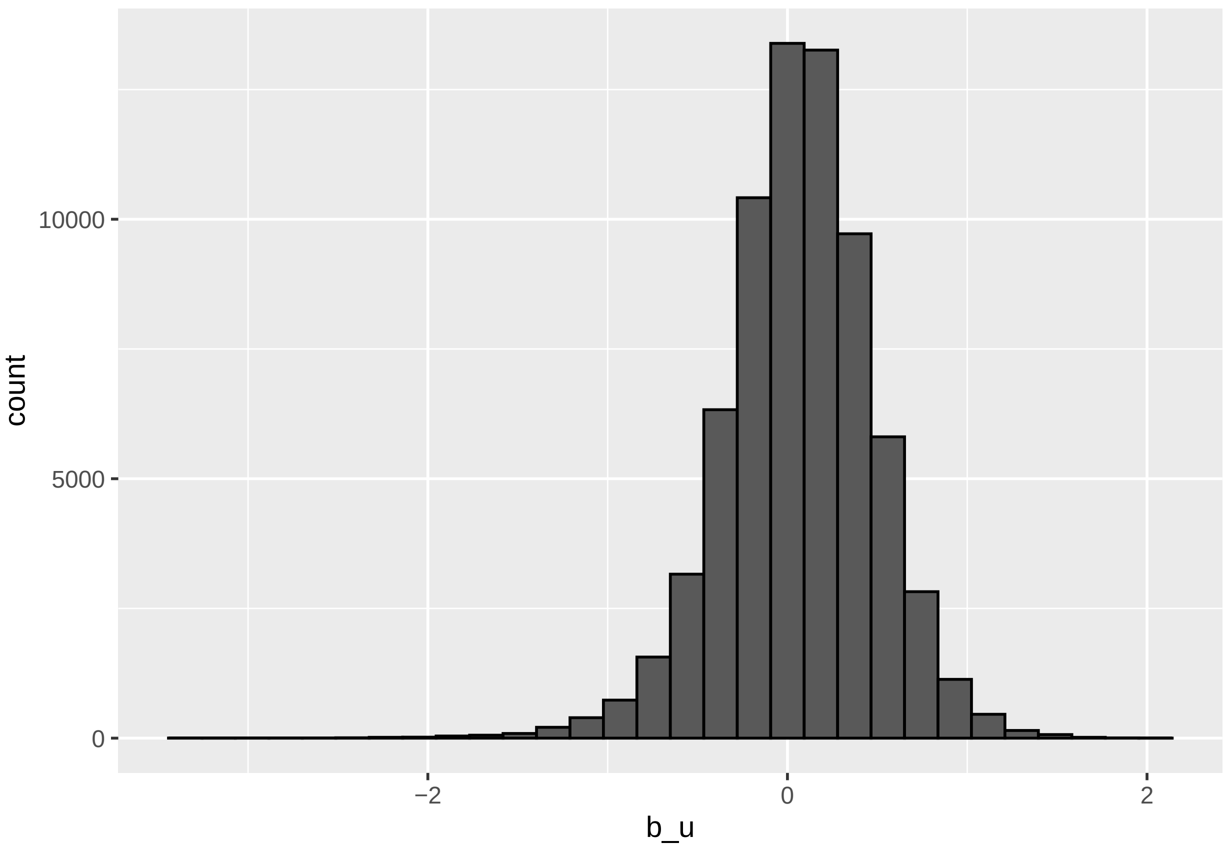
I will also add the user effect to my model. For user effect my model is:

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

RMSE for this model:

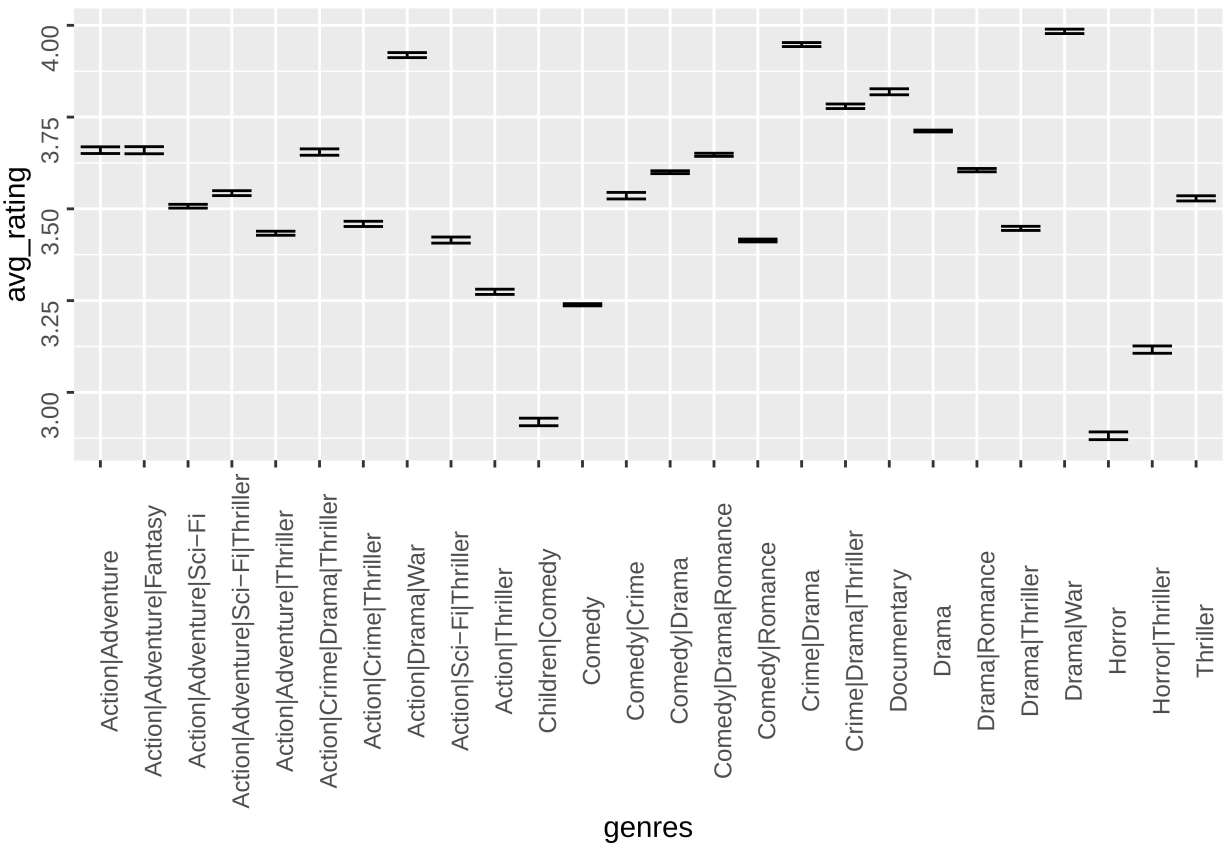
```
## [1] 0.8666408
```

We can also check visually how user to user variability looks like:



Genre Effect

Next I would like to see how genres take part in movie rating. We can visualize genre to genre variability in rating:



As from the plot above we can see, a movie can belong to one or more genres, we need to add up all the genre effects that the movie belongs to . For the genres effect my model is:

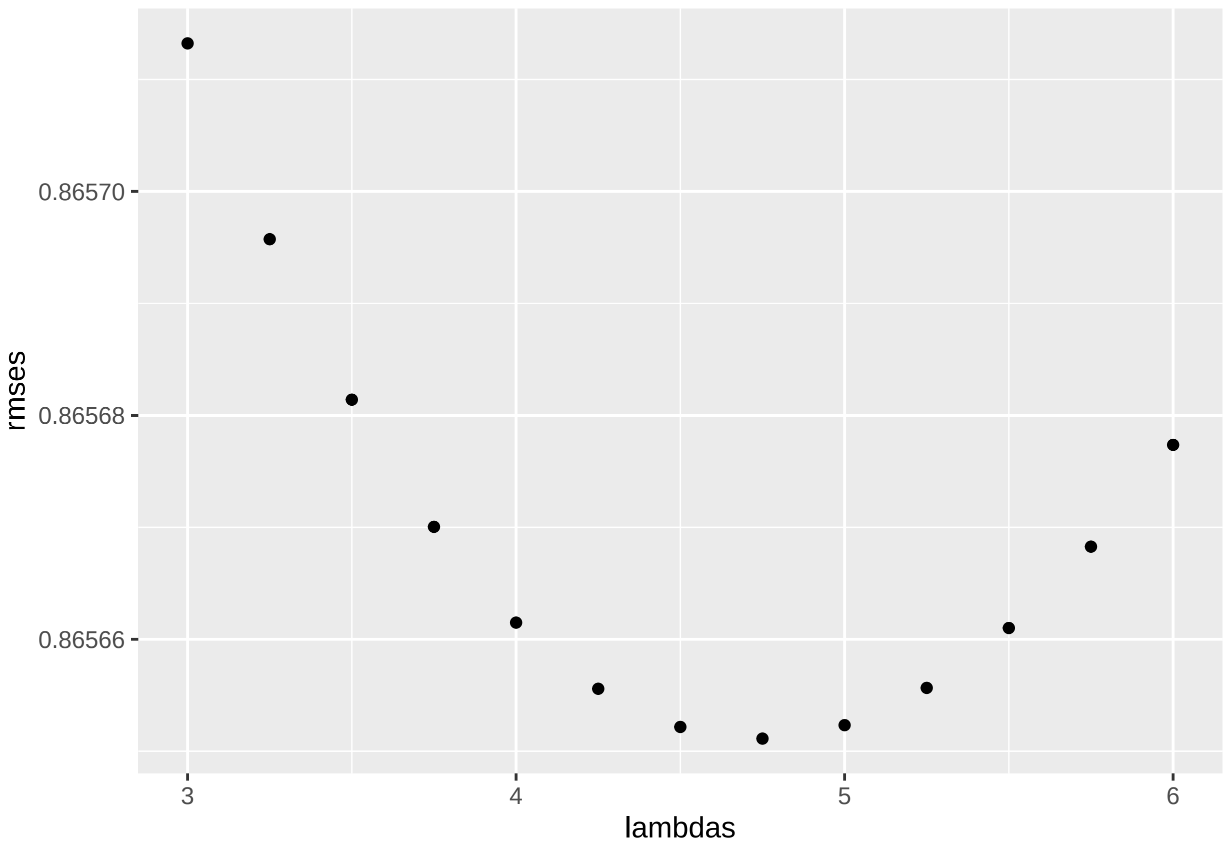
$Y_{u,i} = \mu + b_i + b_u + \sum_{k=1}^{K} X_{u,i} b_k$ with $x_{u,i}^{k} = 1$ if $g_{u,i}$ is genre k.

RMSE for this model:

```
## [1] 0.8662908
```

Regularization

Finally I will use regularization to penalize the genre, user and movie bias and shrink them towards 0 when sample sizes are small. I will use tuning parameter lambda from 3 to 6.



lambda that minimizes RMSE:

```
## [1] 4.75
```

RMSE for regularized movie, user and genre effect on the test set:

```
## [1] 0.8656511
```

Result

From the analysis above we can see that the RMSE keeps getting better as we include user effect, movie effect and genre effect in the model and regularization makes it perform even better. Below is the table that shows RMSEs achieved in different models when testing on the test set (20% of edx set)

```
##                                      Method      RMSE
## 1                                  Just Avg 1.0605613
## 2                              Movie Effect 0.9439868
## 3                               User Effect 0.8666408
## 4                              Genre Effect 0.8662908
## 5 Regularized Movie + User + Genres Effect 0.8656511
```

We can see regularized model with movie, user and genre effect gives us the best RMSE on the test set (20% of edx set), so this is our preferred model.

I will apply the regularized model on the validation set to calculate the final RMSE.

Final RMSE

Final RMSE will be calculated on the validation set. I will use the lambda = 4.75 on the validation set.

Final RMSE on validation set:

```
## [1] 0.8644514
```

Conclusion

In summary, regularization has improved the prediction algorithm and produced the best RMSE (0.8644514) on the validation set. One limitation of this model is that it does not cater for user movie preferences or movie rating pattern. Future work can be done in this area by doing factor analysis using Matrix factorization.