HarvardX PH125.9x - Capstone - Root Mean Square Errors (RMSE) in Movie Rating Predictions

Aditya Wresniyandaka

March 2, 2019

Executive Summary

This document discusses the creation of several models for movie ratings based on the ratings given by users. The data source is based on the 10M version of MovieLens dataset from the GroupLens web site. Exploratory data analysis is performed to understand the data. Several algorithms are built and tested against the validation dataset. The outcomes are the Root Mean Square Error (RMSE) values of each of the algorithms used, and we select the model with the lowest RMSE as the final model.

Analysis Methodology and Approach

Data cleansing and shaping

After downloading the dataset from the GroupLens web site, we performe a preliminary data cleansing and shaping to get the following features: userID, movieId, rating, and timestamp. After that, the data is split into 90% training set (called 'edx'), and 10% testing/validation set (called 'validation').

Exploratory Data Analysis (EDA)

Various exploratory data analysis activities are done further to understand the data, for example the number of rows and columns in the training dataset, the number of unique users and movies, and the top ten movies based on the number of ratings received. Also performed in the EDA is the distribution of the ratings themselves, i.e. what are the ratings used - is it full integers (0 to 5), or does it use half point (0.5, 1, 1.5, etc.).

Building the model and evaluating/calculating the RMSE

The RMSE is established using the simple formula of the root mean squared error between the predicted rating and actual rating. First, we use the Naive method where we calculate the RMSE by comparing it with the overall average rating across all user IDs and movies. Second, we look at the effects of users (user preferences) and movies using the full training dataset and testing/validation dataset from the data cleansing and shapping phase. We calculate the RMSEs based on the movie effect, and based on the movie and user effects. Third, we understand that not every user is actively giving their ratings. For this reason, we filter out users who have only given less than 100 ratings. After that, we run the calculations to obtain the RMSEs. Finally, we compare all the RMSEs from the above steps and select the lowest RMSE.

Exploratory data analysis

Let's do some exploratory data analysis

```
## [1] "The number of rows in the edx dataset: 9000055"
```

```
## [1] "The number of columns in the edx dataset: 6"
```

```
## [1] "The number of zeros in the ratings: 0"
```

```
## [1] "The number of threes in the ratings: 2121240"
```

```
## [1] "The number of movies in the edx dataset: 10677"
## [1] "The number of different users in the edx dataset: 69878"
## [1] "The number of Drama movies: 3910127"
## [1] "The number of Comedy movies: 3540930"
## [1] "The number of Thriller movies: 2325899"
## [1] "The number of Romance movies: 1712100"
```

The genres and the number of ratings received.

genres	count
Drama	**3910127**
Comedy	**3540930**
Action	**2560545**
Thriller	**2325899**
Adventure	**1908892**
Romance	**1712100**
Sci-Fi	**1341183**
Crime	**1327715**
Fantasy	**925637**
Children	**737994**
Horror	**691485**
Mystery	**568332**
War	**511147**
Animation	**467168**
Musical	**433080**
Western	**189394**
Film-Noir	**118541**
Documentary	**93066**
IMAX	**8181**
(no genres listed)	**7**

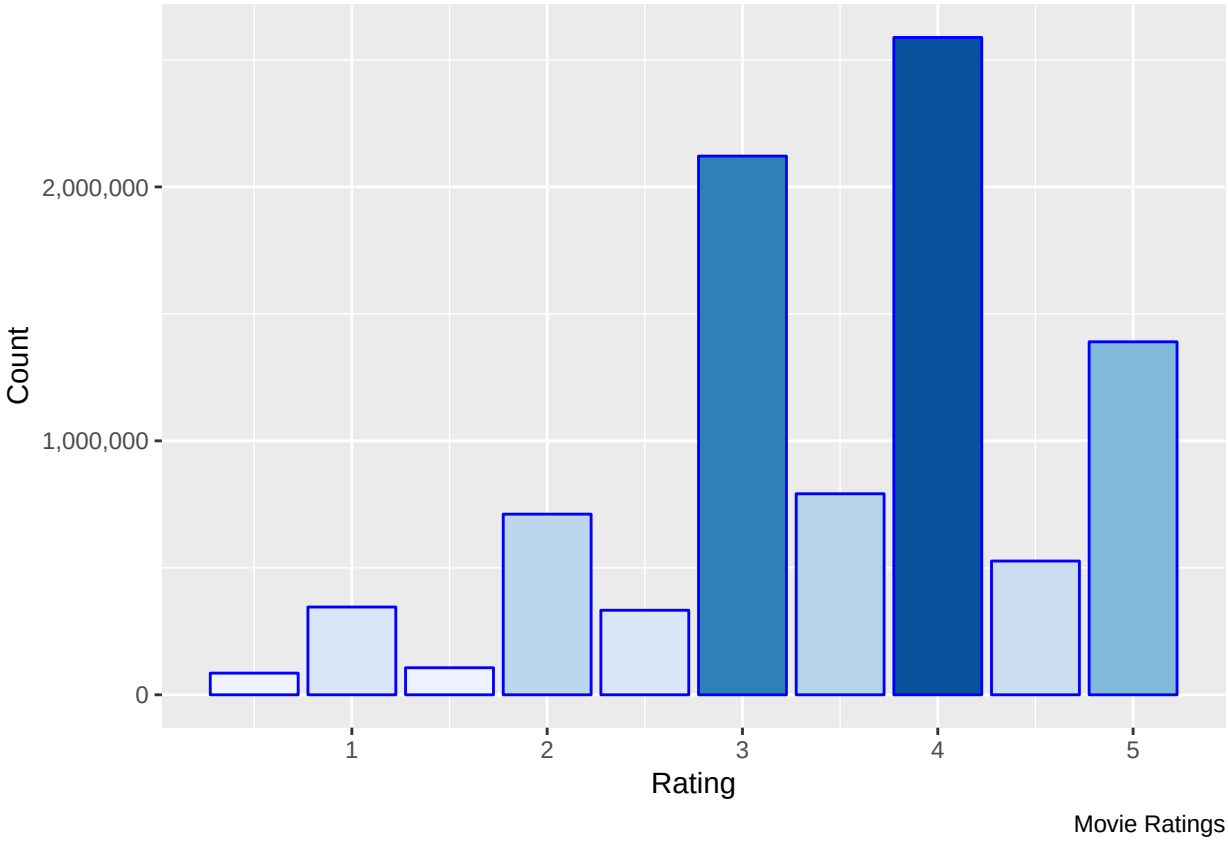
```
## [1] "The movie that has the greatest number of ratings is: Pulp Fiction (1994)"
```

Top 10 movies by the number of ratings received

title	count
Pulp Fiction (1994)	31362
Forrest Gump (1994)	31079
Silence of the Lambs, The (1991)	30382
Jurassic Park (1993)	29360
Shawshank Redemption, The (1994)	28015
Braveheart (1995)	26212
Fugitive, The (1993)	25998
Terminator 2: Judgment Day (1991)	25984
Star Wars: Episode IV - A New Hope (a.k.a. Star Wars) (1977)	25672
Apollo 13 (1995)	24284

The ratings and count of each rating

rating	count
0.5	85374
1.0	345679
1.5	106426
2.0	711422
2.5	333010
3.0	2121240
3.5	791624
4.0	2588430
4.5	526736
5.0	1390114



Building and testing the algorithms

Naive Model

Let us establish the RMSE for the simplest model using the overall average rating across all userIds and movieIds.

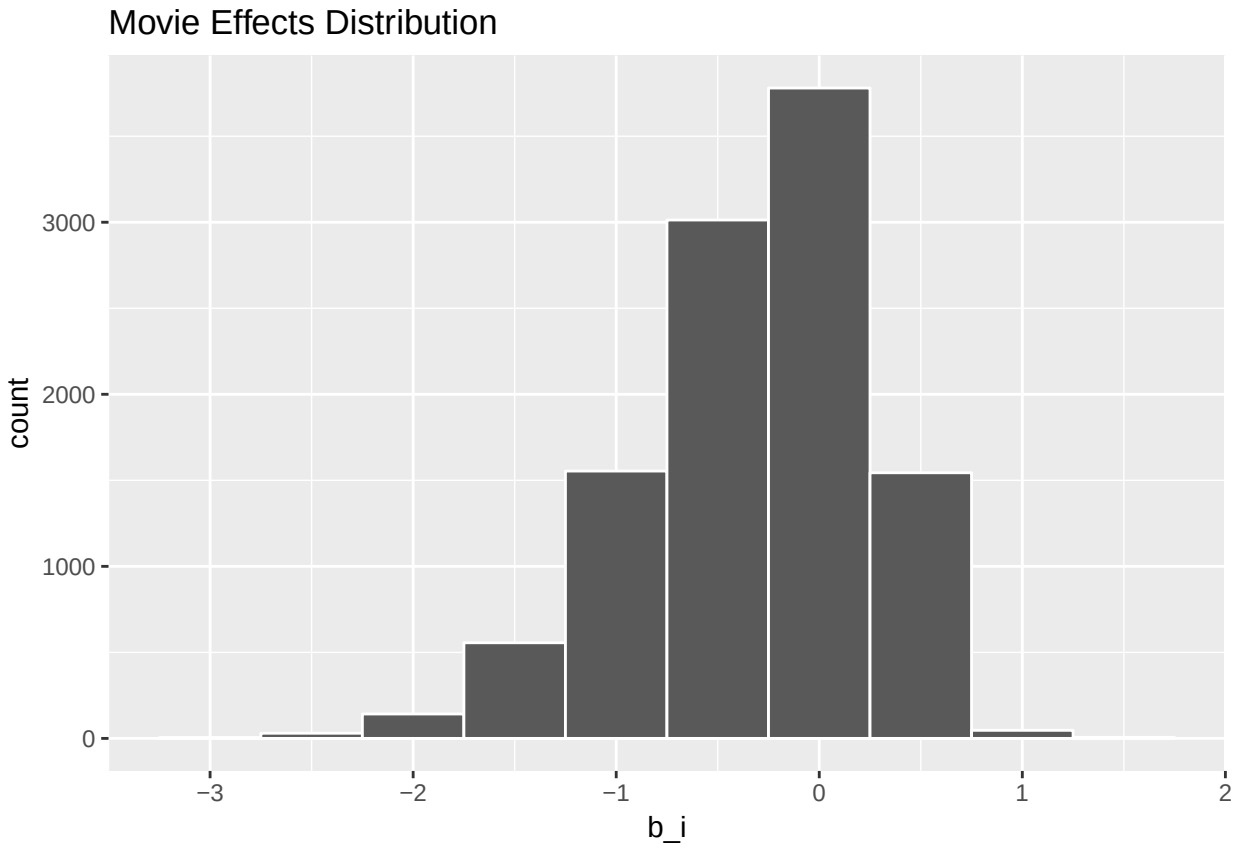
```
## [1] "The RMSE for the Naive Model is: 1.06120181029262"
```

Do we want to use all rows?

Get users who have given at least 100 ratings so we filter out users with less than 100 ratings. Make sure we have both userIds and movieIds in the filtered datasets.

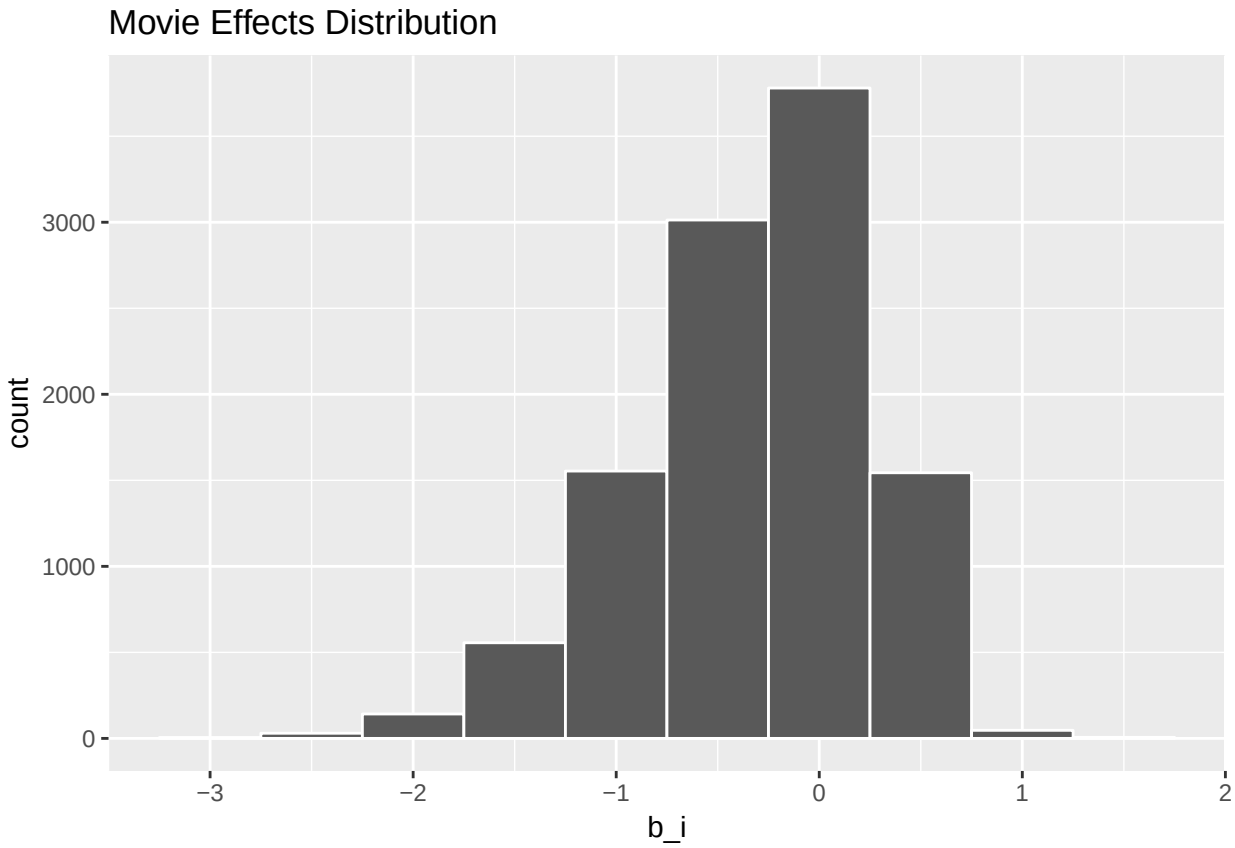
Dataset	Rows
edx original	**9000055**
edx filtered	**6937037**
validation original	**999999**
validation filtered	**768303**

Movie Effect Model



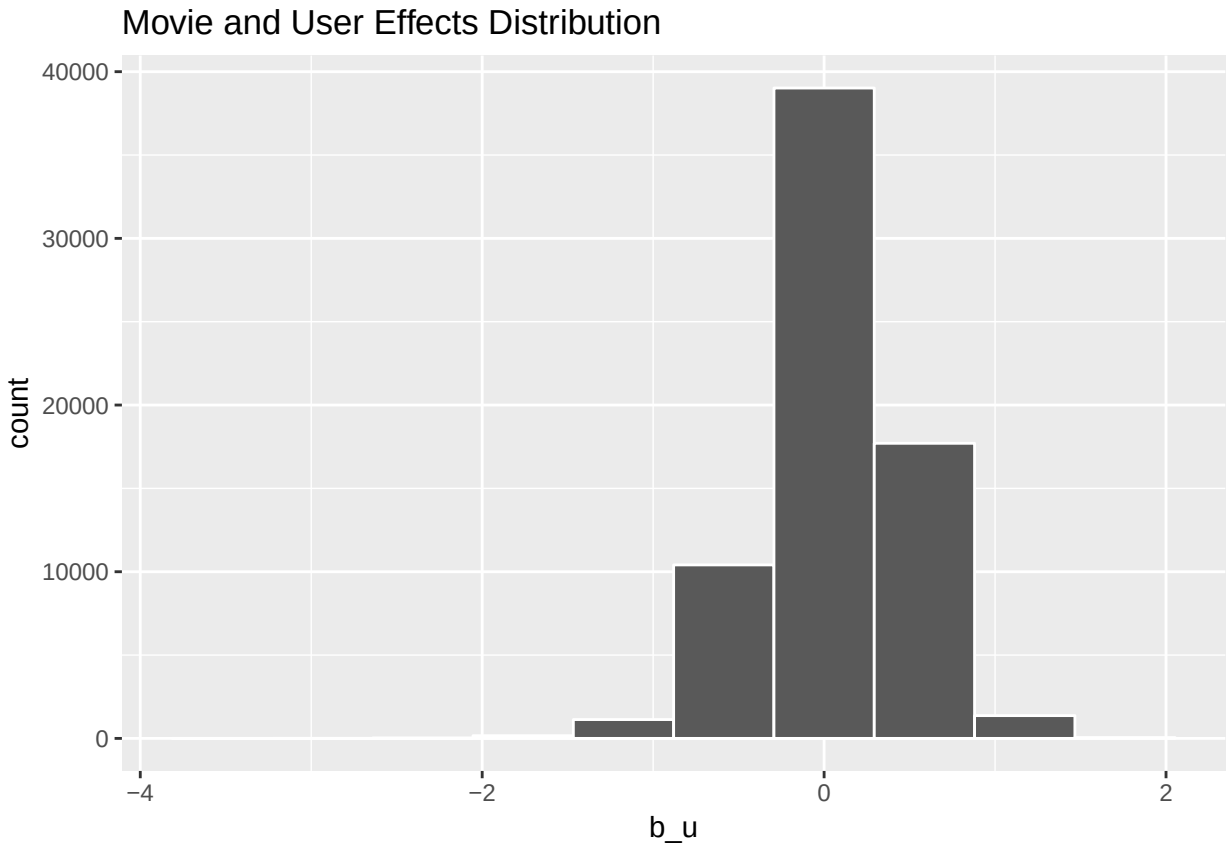

```
## [1] "The RMSE for the Movie Effect Model is: 0.943908662806309"
```

Movie Effect Model with filtered datasets



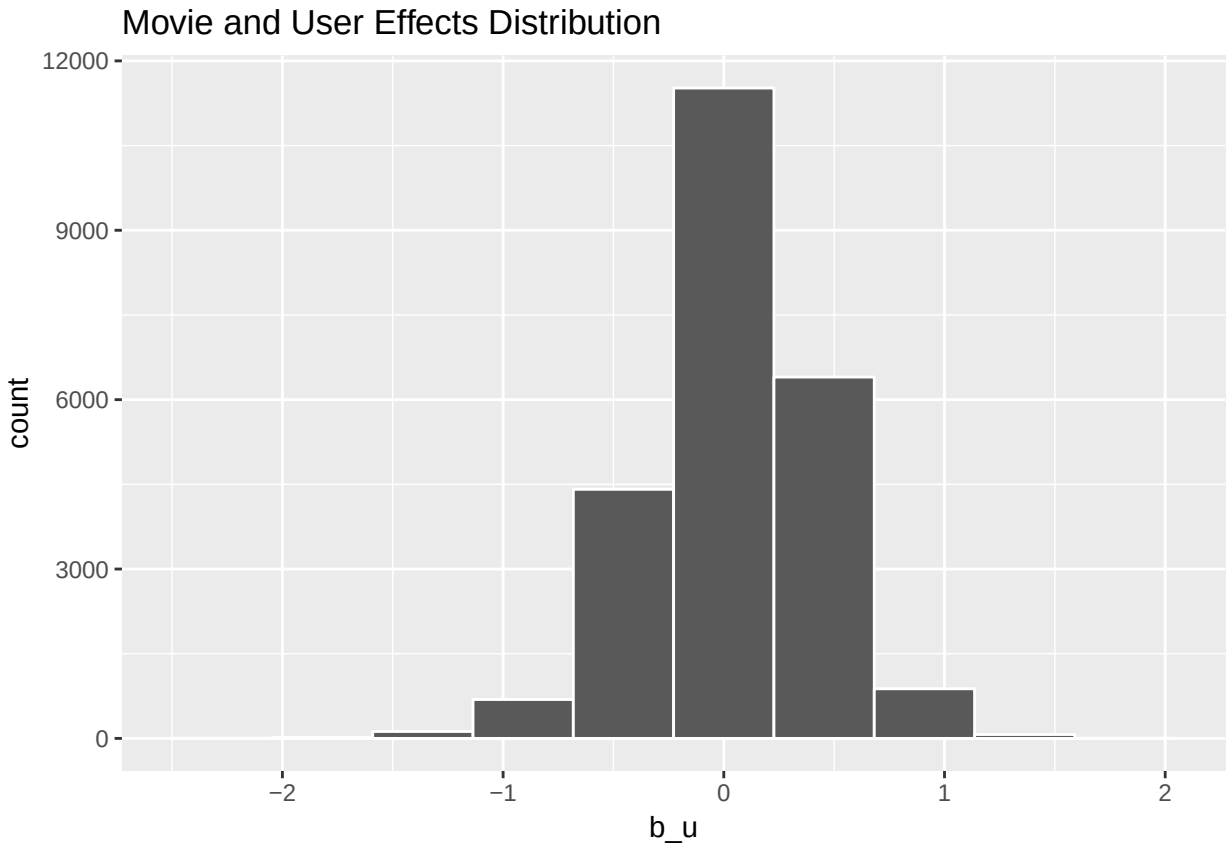

```
## [1] "The RMSE for the Movie Effect Model with filtered datasets is: 0.930215980544284"
```

Movie and User Effects Model




```
## [1] "The RMSE for the Movie and User Effects Model is: 0.865348824577316"
```

Movie and User Effects Model with filtered datasets




```
## [1] "The RMSE for the Movie and User Effects Model with filtered datasets is: 0.849186428036133"
```

RMSE Summary/Results

The results of running the algorithms to obtain various RMSE values can be summarized in the following table.

Method	RMSE
Naive Model	**1.0612018**
Movie Effect Model	**0.9439087**
Movie Effect Model with filtered datasets	**0.9302160**
Movie and User Effects Model	**0.8653488**
Movie and User Effects Model with filtered datasets	**0.8491864**

Conclusion

By including the movie and user effects in the model, we improve the RMSE by 18% (from 1.061 to 0.865) compared to the Naive model. However, by adding a filter to exclude users who have only given ratings less than 100 times, we improve the RMSE further by close to 20% (from 1.061 down to 0.849) compared to the Naive model. The final RMSE we choose is **0.84919** from the combination of using the Movie and User Effects filtered by including users who have given 100 or more ratings.