MovieLens project

Andrea Roetti

June 11, 2020

Contents

Introduction

The MovieLens project goal is to build up a prediction system for movies ratings based on a historical series of records. The dataset we will use as base to create our model is called **edx**, while the dataset where we will measure the efficiency of our model, as if we did not know the given ratings, is called **validation**. In order to get started, we first load some libraries that will be useful for the analysis.

```
library(tidyverse)
library(lattice)
library(caret)
```

We therefore read the two datasets that we have previously already dowloaded as .rds files and saved locally for reasons of size: a file bigger than 100MB is not possible to push to Github, therefore not possible to be kept in the project folders syncronised with it. We know that using complete path rather than relative path is not recommended, in fact below we also provide, as comment the path to the datasets as they were in the "data" folder within our working directory.

```
edx <- readRDS("C:/Users/roetti/Documents/MovieLens/edx.rds")
# edx <- readRDS("data/edx.rds")
validation <- readRDS("C:/Users/roetti/Documents/MovieLens/validation.rds")
# validation <- readRDS("data/validation.rds")
```

To get a first idea of the dataset edx, we here below show the first six records.

```
##   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
```

We see that the dataset has six columns: userId, movieId, rating, timestamp, title and genres. Some of the preliminary and exploratory data anlysis include understanding the number of rows, the number of different users, the number of different movies and the average rating.

```
nrow(edx)
```

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

```
n_distinct(edx$userId)
```

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

```
n_distinct(edx$movieId)
```

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

```
mean(edx$rating)
```

```
## [1] 3.512465
```

In our prediction model, the ultimate goal is to best predict the ratings given in the validation dataset, which has the same structure as the edx. Our criteria to establish how close our predictions fall to the real ratings is the Root Mean Squared Error, or RMSE, a quantity similar to standard deviation: the lower its amount, the more accurate is the prediction. We therefore define here below the function that calculates the RMSE between two strings containing the same number of evaluations (true vs. predicted). We will try, with our predictions, to minimize the value of this fuction.

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

Analysis (part 1)

In this part we summarise the key findings and the approaches used in the Machine Learning course, section 6 "Model Fitting and Recommendation System". We first create a partition of the edx dataset, based on the rating column, into a train set (80% of records) and a test set (20% of records). For the sake of reproducibility, we set the seed to 1 before doing it.

```
set.seed(1, sample.kind = "Rounding")
test_index <- createDataPartition(y = edx$rating, times = 1,
                                  p = 0.2, list = FALSE)
train_set <- edx[-test_index,]
test_set <- edx[test_index,]
```

We also want to make sure that we don't include users and movies in the test set that do not appear in the training set, therfore we run the following code.

```
test_set <- test_set %>%
     semi_join(train_set, by = "movieId") %>%
     semi_join(train_set, by = "userId")
```

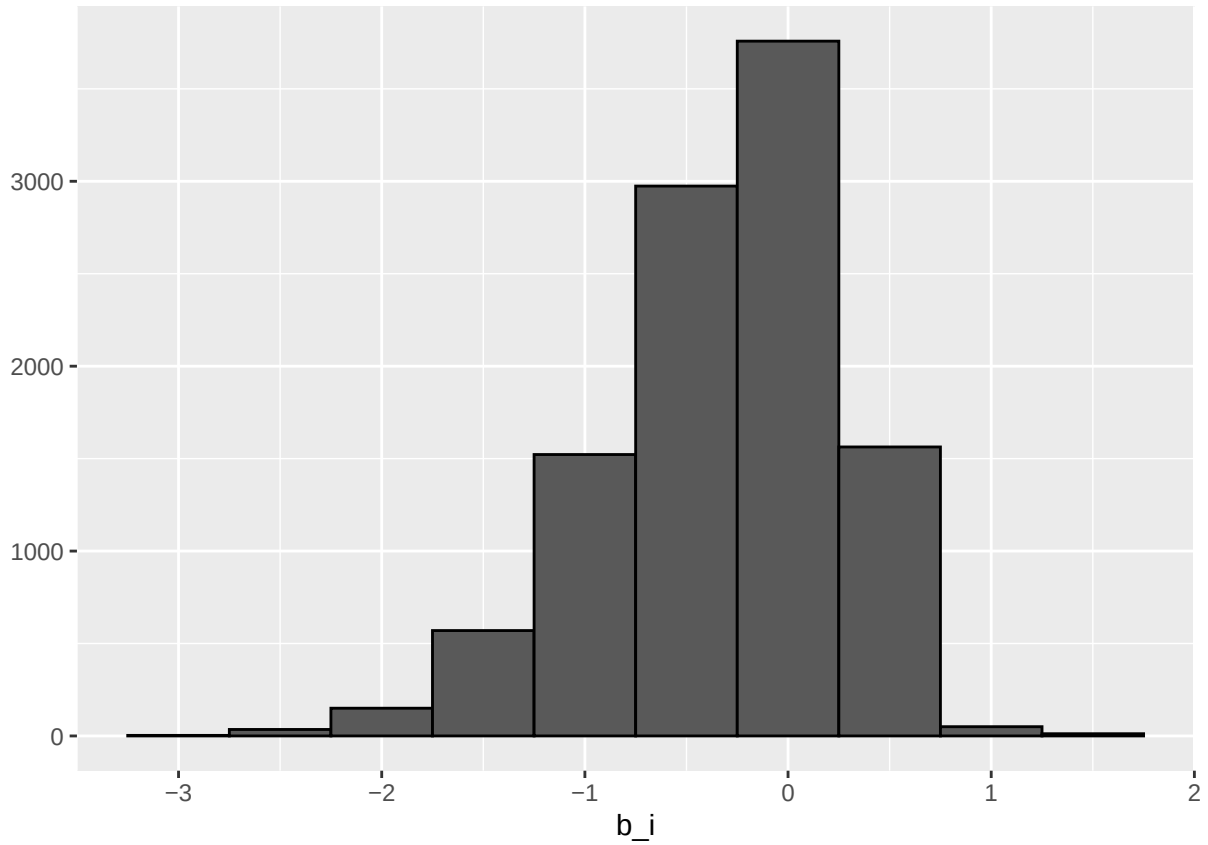
Knowing that the average rating is the constant that minimizes the RMSE, we take, as a first benchmark of our model built on the train set, the average rating in the train set, assigned to mu. We then test its effectivness on the test set and report the result in a table to keep track of our improvements.

```
mu <- mean(train_set$rating)
first_rmse <- RMSE(test_set$rating, mu)
rmse_results <- tibble(method = "Just the average", RMSE = first_rmse)
rmse_results %>% knitr::kable()
```

method	RMSE
Just the average	1.059904

The result, almost 1.06 is quite high. To improve the result, we first think that different movies have different average ratings. We call this a bias, namely a factor, the movie ID, that impacts our predicted rating. To check our guess, we subtract the average rating from the movie average, so that our b_i, bias factors due to the movie ID, are positive only if the movie has an average rating above the overall average mu, negative if below. Then we plot an histogram.

```
b_i <- train_set %>%
    group_by(movieId) %>%
    summarize(b_i = sum(rating - mu)/(n()))
b_i %>% qplot(b_i, geom ="histogram", bins = 10, data = ., color = I("black"))
```



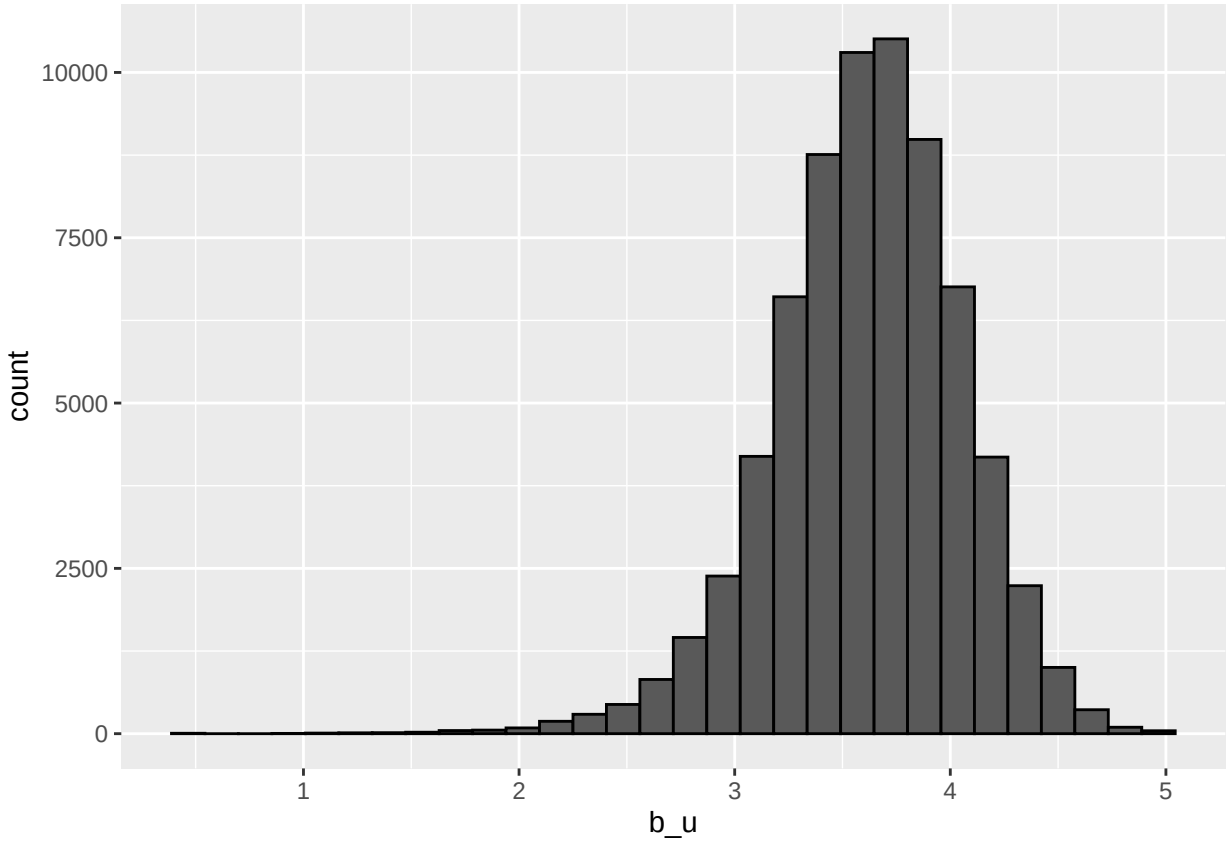
Our guess is confirmed, we therefore include the factors b_i in our prediction system and test again the effectivness on the test set. We call this predictions second_pred, since the first ones simply coincided with mu and we report the new rmse in the same table as before.

```
second_pred <- test_set %>%
     left_join(b_i, by='movieId') %>%
     mutate(pred = mu + b_i) %>% .$pred
second_rmse <- RMSE(test_set$rating, second_pred)
rmse_results <- bind_rows(rmse_results,
                          tibble(method="Movie Effect Model",
                                 RMSE = second_rmse))
rmse_results %>% knitr::kable()
```

method	RMSE
Just the average	1.0599043
Movie Effect Model	0.9437429

The result has quite improved. However, we suspect that a possible bias could come also from the userId, since different users reasonably give different average ratings. We try to verify this by plotting an histogram of the average ratings for users that have rated at least 100 movies.

```
train_set %>%
     group_by(userId) %>%
     summarize(b_u = mean(rating)) %>%
     filter(n()>=100) %>%
     ggplot(aes(b_u)) +
     geom_histogram(bins = 30, color = "black")
```



The plot confirms our suspect, therefore we include the bias factors b_u linked to the userId in our prediction system. As before, we calculate the b_u from the training set, we test the model and we update the table with the new RMSE.

```
b_u <- train_set %>%
     left_join(b_i, by="movieId") %>%
     group_by(userId) %>%
     summarize(b_u = sum(rating - b_i - mu)/(n()))
third_pred <- test_set %>%
     left_join(b_i, by='movieId') %>%
     left_join(b_u, by='userId') %>%
```

```
    mutate(pred = mu + b_i + b_u) %>% .$pred
third_rmse <- RMSE(test_set$rating, third_pred)
rmse_results <- bind_rows(rmse_results,
                          tibble(method="Movie + User Effects Model",
                                     RMSE = third_rmse))
rmse_results %>% knitr::kable()
```

method	RMSE
Just the average	1.0599043
Movie Effect Model	0.9437429
Movie + User Effects Model	0.8659320