Final Project

Liam Montero Guillemi

2025-08-08

Contents

Introduction

What will the project do?

In this project, we want to create a movie recommendation system, in which we will use RMSE as a measure of loss.

Why is a good recommendation system important?

A good recommendation system is crucial because it allows new users to discover personalized movies from the start, increasing their satisfaction with the platform. It facilitates the discovery of content beyond what is popular, exposing users to options that truly interest them. This not only improves the user experience but also plays a fundamental role in long-term retention by offering continuous value through relevant suggestions. Ultimately, an effective system is key to the platform's success and competitiveness.

How does it benefit businesses and users?

A good recommendation system benefits both businesses and users. For businesses, it boosts retention, increases revenue, provides valuable user data, and offers a competitive advantage. For users, it facilitates personalized movie discovery, reduces information overload, exposes them to new content, and improves the overall platform experience, saving them time and increasing satisfaction. In essence, it creates a mutually beneficial relationship by connecting users with the content they love and businesses with more engaged audiences.

Creating the dataset

This part of the project was provided by the Edx platform, below we will explain step by step each line of it. The following lines of code install the tidyverse and caret packages if they aren't already installed. They are then loaded using the library() function.

```
if(!require(tidyverse)) install.packages("tidyverse", repos = "http://cran.us.r-project.org")
if(!require(caret)) install.packages("caret", repos = "http://cran.us.r-project.org")
library(tidyverse)
library(caret)
```

Then we have the following 3 lines of code, which perform the following actions: 1) download the 10 million data document from the MovieLens dataset 2) extract the ratings data from the downloaded file 3) extract the movie name data

```
dl <- "ml-10M100K.zip"
if(!file.exists(dl))
  download.file("https://files.grouplens.org/datasets/movielens/ml-10m.zip", dl)

ratings_file <- "ml-10M100K/ratings.dat"
if(!file.exists(ratings_file))
  unzip(dl, ratings_file)

movies_file <- "ml-10M100K/movies.dat"
if(!file.exists(movies_file))
  unzip(dl, movies_file)
```

After that, the data is converted into a format that is easy to work with, in this case it is a data frame, everything that is inside the ratings_file file is read, then each of its rows is divided into columns using the separator "::"

```
ratings <- as.data.frame(str_split(read_lines(ratings_file), fixed("::"), simplify = TRUE),
                         stringsAsFactors = FALSE)
```

then each of these columns is named as follows:

```
colnames(ratings) <- c("userId", "movieId", "rating", "timestamp")
```

The data frame is configured so that each of the columns has the data in the desired format:

```
ratings <- ratings %>%
  mutate(userId = as.integer(userId),
         movieId = as.integer(movieId),
         rating = as.numeric(rating),
         timestamp = as.integer(timestamp))
```

We do the same thing as we did with the ratings_file, but this time with the movies_file, that is, we convert the data to an easy-to-work format, name its columns, and configure the data to a desired format:

```
movies <- as.data.frame(str_split(read_lines(movies_file), fixed("::"), simplify = TRUE),
                        stringsAsFactors = FALSE)

colnames(movies) <- c("movieId", "title", "genres")

movies <- movies %>%
  mutate(movieId = as.integer(movieId))
```

The 2 data frames created previously are joined using the movieId column:

```
movielens <- left_join(ratings, movies, by = "movieId")
```

Now, we proceed to divide the data set into a training set and a test set:

```
set.seed(1, sample.kind="Rounding") # if using R 3.6 or later
test_index <- createDataPartition(y = movielens$rating, times = 1, p = 0.1, list = FALSE)
edx <- movielens[-test_index,]
temp <- movielens[test_index,]
```

We make sure that all data in the test set is also in the training set as follows:

```
final_holdout_test <- temp %>%
  semi_join(edx, by = "movieId") %>%
  semi_join(edx, by = "userId")
```

The rows that were removed from the test set are added to the training set so that this information is not lost:

```
removed <- anti_join(temp, final_holdout_test)
edx <- rbind(edx, removed)
```

Finally, all objects created to shape the training and test sets but no longer used are deleted from the workspace. This is done to free up memory space and allow for more comfortable work.

```
rm(dl, ratings, movies, test_index, temp, movielens, removed)
```

Data Exploration

In this section an overview over the edx dataset is made. In Table 1 and Table 2 the number of rows and the structure of the data in the first 10 rows can be seen.

```
##      Rows
## 1 9000055
```

```
##    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)
## 8       1     356      5 838983653                       Forrest Gump (1994)
## 9       1     362      5 838984885                   Jungle Book, The (1994)
## 10      1     364      5 838983707                     Lion King, The (1994)
## 11      1     370      5 838984596 Naked Gun 33 1/3: The Final Insult (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
## 8                   Comedy|Drama|Romance|War
## 9                 Adventure|Children|Romance
## 10 Adventure|Animation|Children|Drama|Musical
## 11                            Action|Comedy
```

Ratings

In the data, movie ratings are included. Possible ratings were in the range between 0.5 and 5. Hereby, whole numbers seem to be preferred. The distribution of the number of ratings on the different rating values can be seen in Figure 1. In Table 3, the top 10 films with the most ratings are listed.

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

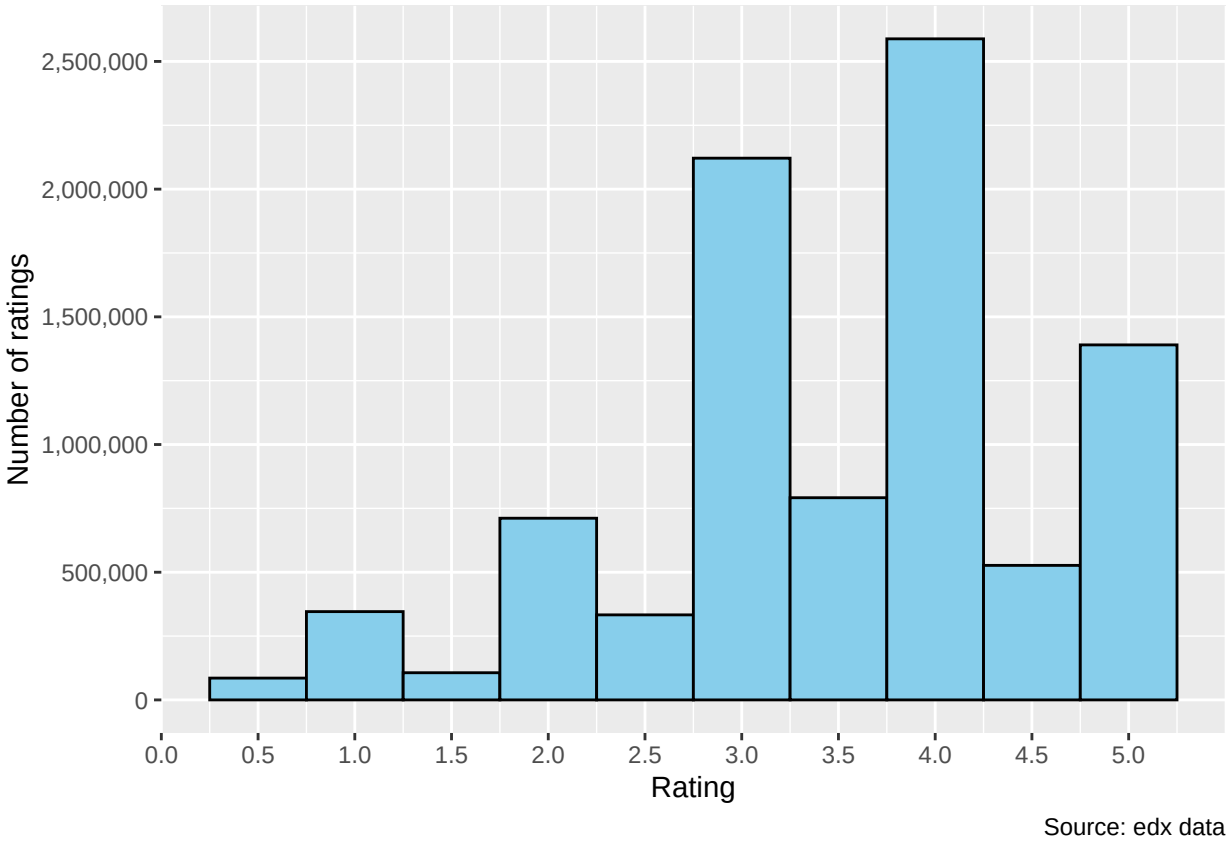

Figure 1: Distribution of film ratings

The dataset includes ratings from 1995 to 2009. During that time, the number of ratings made changes. In Figure 2, the number of ratings are summed up per week and plotted by time. It can be seen that before 2000, the number of ratings varied a lot. After 2000, the number of ratings reached a certain level with some peaks, decreased again but stayed at a level of about 22,000 ratings per week till 2009. The peaks may possibly come from a blockbuster movie coming out at that time.

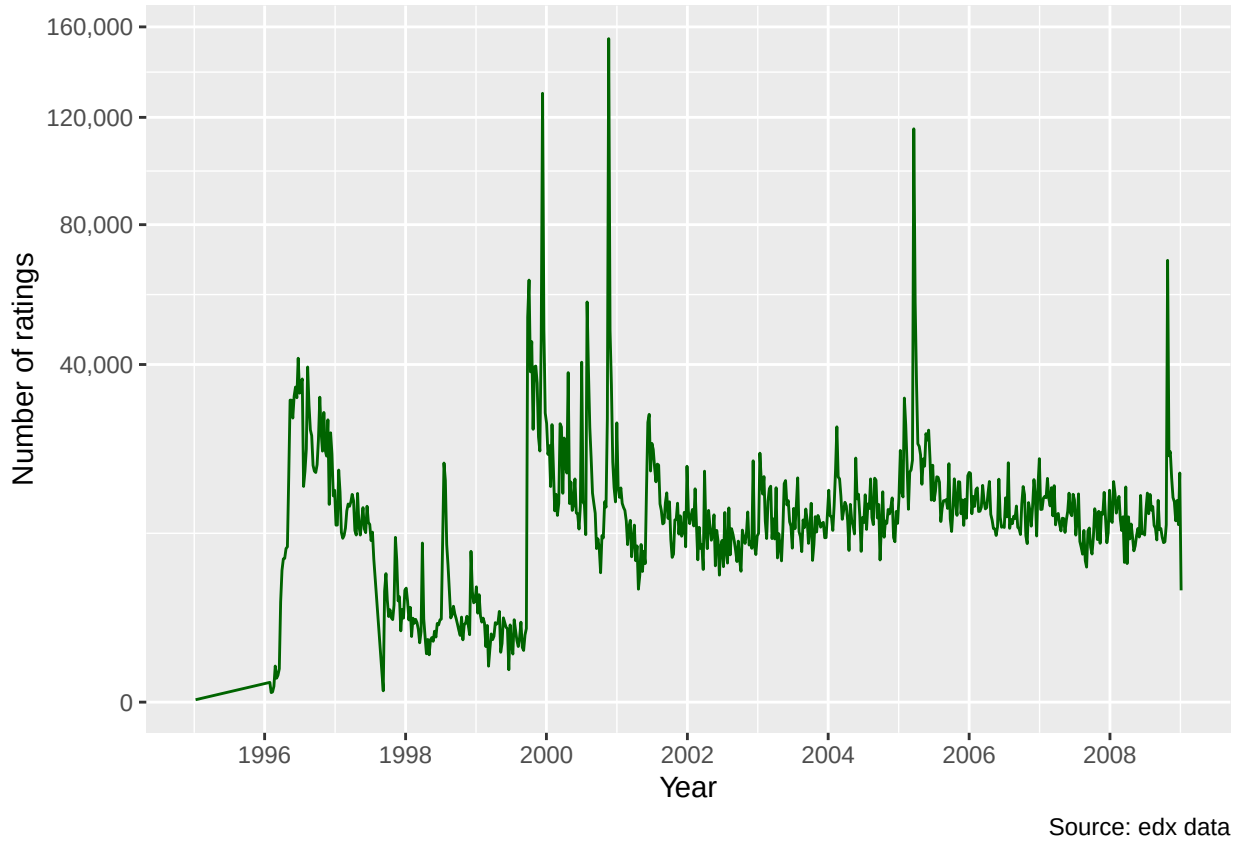

Figure 2: Number of ratings over time (weekly)

Related observations

The next two diagrams in Figure 4 show the average rating vs. number of users, respectively, the average number of ratings per user. Most of the users have an average rating overall of 3.5 - 3.8. This seems to follow a Gaussian distribution. On the right, it can be seen that there are some "power users" (around 200 - 400) who have several thousand ratings. However, most of the users only made a small number of ratings.

Time dependent observations

In the dataset, there are time dimensions that should be mentioned. The first one is the time of rating. Possibly, in some time period or time of the year, the average ratings are different. So, there could possibly be a time effect on the rating value. The second time dimension is the release year of the specific movie. This value can be extracted from the title in the dataset. If the average rating per release year of the film is plotted in a diagram, it can be seen that older films tend to have a higher rating than newer films. So, both effects could play a role in developing the algorithm to predict the ratings. The two diagrams can be seen in Figure 5.

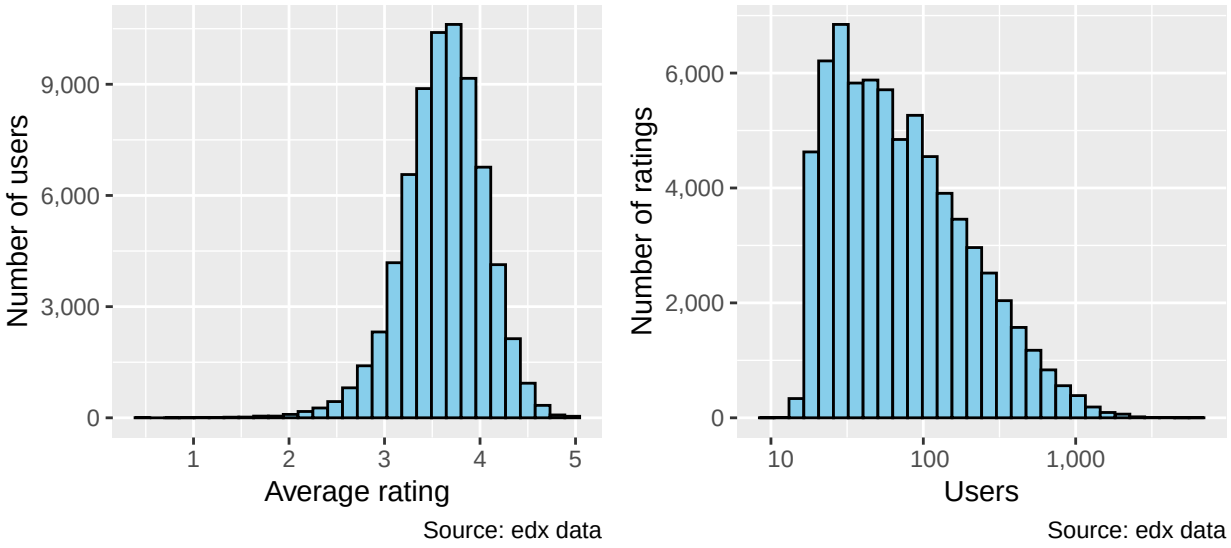

Figure 3: Average rating per user (left) and average number of ratings per user (right)

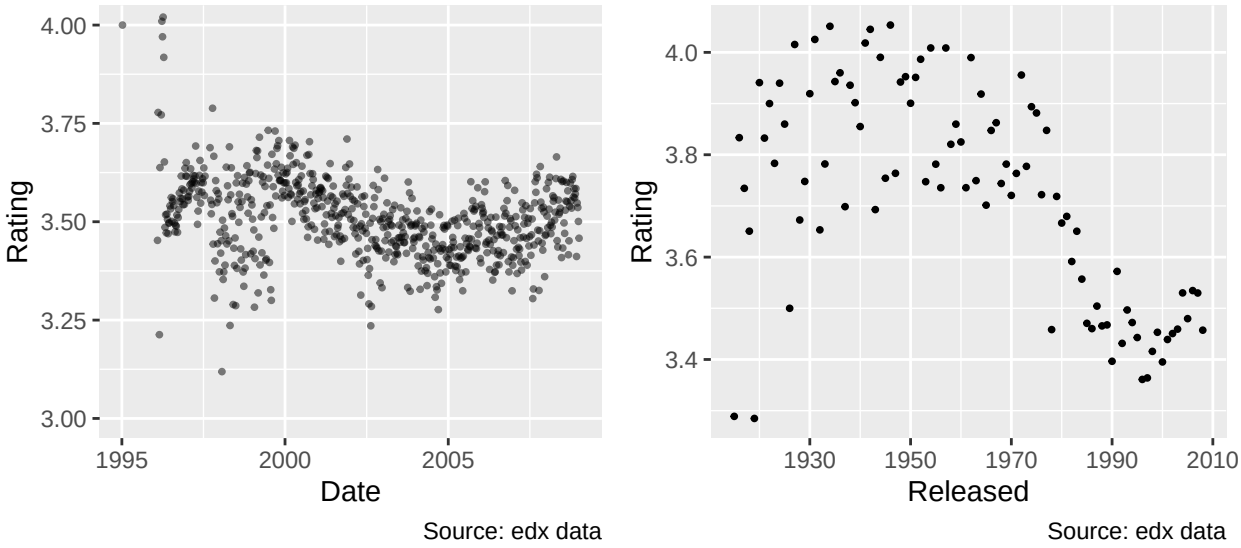

Figure 4: Date vs. weekly averaged rating (left) and release year vs. averaged rating (right)

Creating the model

How did I find which model to use?

I'd like to start by saying that I decided to use the matrix factorization method because it offers the best results on large datasets. I decided to use the recosystem package and not the recommenderlab package because the recosystem package is optimized for large datasets, while the recommenderlab package is not. I learned this through trial and error. With the recommenderlab package, my computer never managed to finish the model; it always shut down before finishing it. I spent many hours running calculations. I think an Intel Core i9-13980HX isn't a bad processor, and paired with 32GB of RAM, it's not bad at all. Therefore, I decided to change my approach and use recosystem, which is designed to work with large datasets. It can perform parallel operations using the CPU, which greatly shortens processing times. It can load the training and prediction sets from the hard drive as a file, which means it doesn't overload the RAM too much. Although it also has the option to load them from the R workspace.

Implementation of the model

The first thing I did was install and load recosystem

```
if(!require(recosystem)) install.packages("recosystem")
if(!require(parallel)) install.packages("parallel")
library(recosystem)
library(parallel)
```

Then, we adjust the training set with only the parameters we are going to use, this is not absolutely necessary, but I like to do it to be more organized.

```
train_set <- edx %>% select(userId, movieId, rating) %>% data.frame()
```

We convert the training set into a format that the recosystem ecosystem can interpret, for this we use the data_memory() function, which uses an object that is in the R workspace and converts it into a DataSource object, which is the type of object used to train and predict models with the recosystem library.

```
train_data <- data_memory(train_set[, 1], train_set[, 2], rating = train_set[, 3])
```

creation of a RecoSys object, which will contain all the parameters, training and predictions.

```
r <- Reco()
```

Next, a hyperparameter adjustment is made, this is carried out by the r$tune function, which has a large list of many parameters to adjust, the vast majority of the parameters I chose were those that came by default, this because with chatGPT I did a search for the best parameters for large data sets and this gave me an answer that was very similar to the parameters that came by default. In the parameter nthread = detectCores() - 1, I put it like this because what this parameter does is use all the cores of your CPU except 1, and the parameter nbin = 32, is because it always has to be greater than the nthread parameter, so, since my CPU has 32 cores, nthread = 31 and nbin = 32.

This function performs cross-validation, which has 5 folds by default (which can also be adjusted), and runs all possible combinations of all parameters to find the best combination of them that results in a lower RMSE, if I wanted to evaluate using another loss function, I could also do so by adjusting the loss parameter, which has the RMSE by default.

```
best_tune <- r$tune(train_data, opts = list(dim = c(10L, 20L, 30L),
                                costp_l1 = c(0, 0.1),
                                costp_l2 = c(0.01, 0.1),
                                costq_l1 = c(0, 0.1),
                                costq_l2 = c(0.01, 0.1),
                                lrate    = c(0.01, 0.1),
                                nthread  = detectCores() - 1,
                                nbin     = detectCores())
)
```

Here I proceed to train the recommendation model with r$train, using the best parameters found with the r$tune function

```
final_model <- r$train(train_data, opts = best_tune$min)
```

Then, we proceed to prepare the evaluation data set, just as we did with the test data set, but this time we do not include the ratings, because that is what I am supposed to predict.

```
test_data <- data_memory(final_holdout_test[, 1], final_holdout_test[, 2])
```

We make predictions with r$predict and the object obtained is a vector, which contains all the ratings predicted by the model for the users.

```
pred2 = r$predict(test_data, out_memory())
```

Finally, we proceed to calculate the RMSE, to see the accuracy of our model, we do this using the vector of real values of the evaluation set and the vector obtained from the model's prediction on the evaluation set. General form of the least squares loss formula:

$RMSE < -sqrt(mean((true.values - predict.values)^2))$

```
RMSE <- sqrt(mean((final_holdout_test[, 3] - pred2)^2))
RMSE
```

```
## [1] 0.782828
```

Conclusions

This project has enabled the development of a movie recommendation system based on matrix factorization techniques, demonstrating the efficiency of machine learning methods applied to large datasets. The most important points are highlighted below:

• **Data Preprocessing and Organization:** A rigorous extraction and transformation process was implemented for the Movielens dataset, ensuring proper data structuring from files with millions of records. The division into training and test sets, along with data integrity validation, was essential to build a solid foundation on which to train and evaluate the model.

• **Choice of Tools and Methodology:** The decision to use the recosystem package instead of recommenderlab was based on the need to handle a large volume of data and on computational efficiency. Recosystem enabled hyperparameter optimization through cross-validation and leveraged parallelization, resulting in a significant reduction in processing times and more efficient management of hardware resources.

• **Reflections and Future Directions:** This project demonstrates the importance of integrating data science techniques with robust optimization methodologies for handling large volumes of data. It also lays the groundwork for future improvements, such as the incorporation of new algorithms, further hyperparameter optimization, or adaptation of the system to other domains. The experience gained demonstrates that the combination of good preprocessing, the choice of appropriate tools, and an iterative approach to validation can lead to highly effective solutions in real-world environments.