

Capstone Movielens Report

Azamat Kurbanayev

2025-05-16

Contents

1	Introduction / Overview / Executive Summary	1
1.1	Datasets Overview	3
2	Methods / Analysis	15
2.1	Defining Logging and Time Measuring Helper Functions	15
2.2	Preparing train and test datasets	17
2.3	Overall Mean Rating (Naive) Model	27
2.4	User Effect Model	32
2.5	User+Movie Effect (UME) Model	36
2.6	User+Movie+Genre Effect (UMGE) Model	46
2.7	User+Movie+Genre+Year Effect (UMGYE) Model	49
2.8	User+Movie+Genre+Year+Day Effect (UMGYDE) Model	51
3	Conclusion	53
4	Appendix	54
4.1	Data Helper Functions	54
4.2	Regularization: Common Helper Functions	56
4.3	Support Functions	69

1 Introduction / Overview / Executive Summary

The goal of the project is to build a Recommendation System using a [10M version of the MovieLens dataset](#). Following the [Netflix Grand Prize Contest](#) requirements, we will evaluate the *Root Mean Square Error (RMSE)* score, which, as shown in [Section 23.2 Loss function](#) of the *Course Textbook*, is defined as:

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{i,j}^N (y_{i,j} - \hat{y}_{i,j})^2}$$

with N being the number of user/movie combinations for which we make predictions and the sum occurring over all these combinations[1].

Our goal is to achieve a value of less than 0.86490 (compare with the *Netflix Grand Prize* requirement: of at least 0.8563[2]).

1.1 Datasets Overview

To start with we have to generate two datasets derived from the *MovieLens* one mentioned above:

- **edx**: we use it to develop and train our algorithms;
- **final_holdout_test**: according to the course requirements, we use it exclusively to evaluate the *RMSE* of our final algorithm.

For this purpose the following package has been developed by the author of this report: `edx.capstone.movieLens.data`. The source code of the package is available [on GitHub](#)[3].

Let's install the development version of this package from the GitHub repository and attach the correspondent library to the global environment:

```
if(!require(edx.capstone.movieLens.data)) pak::pak("AzKurban-edX-DS/edx.capstone.movieLens.data")

library(edx.capstone.movieLens.data)
edx <- edx.capstone.movieLens.data::edx
final_holdout_test <- edx.capstone.movieLens.data::final_holdout_test
```

Now, we have the datasets listed above:

```
summary(edx)
```

```
##      userId        movieId       rating      timestamp        title      genres
##  Min.   :    1   Min.   :    1   Min.   :0.500   Min.   :7.897e+08   Length:9000055   Length:90000
##  1st Qu.:18124  1st Qu.:  648  1st Qu.:3.000   1st Qu.:9.468e+08   Class  :character  Class  :chara
##  Median :35738  Median : 1834  Median :4.000   Median :1.035e+09   Mode   :character  Mode   :chara
##  Mean   :35870  Mean   : 4122  Mean   :3.512   Mean   :1.033e+09
##  3rd Qu.:53607  3rd Qu.: 3626  3rd Qu.:4.000   3rd Qu.:1.127e+09
##  Max.   :71567  Max.   :65133  Max.   :5.000   Max.   :1.231e+09
```

```
summary(final_holdout_test)
```

```
##      userId        movieId       rating      timestamp        title      genres
##  Min.   :    1   Min.   :    1   Min.   :0.500   Min.   :7.897e+08   Length:999999   Length:999999
##  1st Qu.:18096  1st Qu.:  648  1st Qu.:3.000   1st Qu.:9.467e+08   Class  :character  Class  :chara
##  Median :35768  Median : 1827  Median :4.000   Median :1.035e+09   Mode   :character  Mode   :chara
##  Mean   :35870  Mean   : 4108  Mean   :3.512   Mean   :1.033e+09
##  3rd Qu.:53621  3rd Qu.: 3624  3rd Qu.:4.000   3rd Qu.:1.127e+09
##  Max.   :71567  Max.   :65133  Max.   :5.000   Max.   :1.231e+09
```

1.1.1 edx Dataset

Let's look into the details of the `edx` dataset:

```
str(edx)
```

```
## 'data.frame': 9000055 obs. of 6 variables:  
## $ userId    : int 1 1 1 1 1 1 1 1 1 1 ...  
## $ movieId   : int 122 185 292 316 329 355 356 362 364 370 ...  
## $ rating    : num 5 5 5 5 5 5 5 5 5 5 ...  
## $ timestamp: int 838985046 838983525 838983421 838983392 838983392 838984474 838983653 838984885 838984885 838984885 ...  
## $ title     : chr "Boomerang (1992)" "Net, The (1995)" "Outbreak (1995)" "Stargate (1994)" ...  
## $ genres    : chr "Comedy|Romance" "Action|Crime|Thriller" "Action|Drama|Sci-Fi|Thriller" "Action|Drama|Sci-Fi|Thriller" "Action|Drama|Sci-Fi|Thriller" ...
```

Note that we have 9000055 rows and six columns in there:

```
dim_edx <- dim(edx)  
print(dim_edx)
```

```
## [1] 9000055      6
```

First, let's note that we have 10677 different movies:

```
n_movies <- n_distinct(edx$movieId)  
print(n_movies)
```

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

and 69878 different users in the dataset:

```
n_users <- n_distinct(edx$userId)  
print(n_users)
```

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

Now, note the expressions below which confirm the fact explained in [Section 23.1.1 Movielens data](#) of the *Course Textbook*[4] that not every user rated every movie:

```
max_possible_ratings <- n_movies*n_users  
sprintf("Maximum possible ratings: %s", max_possible_ratings)
```

```
## [1] "Maximum possible ratings: 746087406"
```

```
sprintf("Rows in `edx` dataset: %s", dim_edx[1])
```

```
## [1] "Rows in 'edx' dataset: 9000055"
```

```
sprintf("Not every movie was rated: %s", max_possible_ratings > dim_edx[1])
```

```
## [1] "Not every movie was rated: TRUE"
```

As also explained in that section, we can think of these data as a very large matrix, with users on the rows and movies on the columns, with many empty cells. Therefore, we can think of a recommendation system as filling in the NAs in the dataset for the movies that some or all the users do not rate. A sample from the edx data below illustrates this idea[5]:

```
keep <- edx |>
  dplyr::count(movieId) |>
  top_n(4, n) |>
  pull(movieId)

tab <- edx |>
  filter(movieId %in% keep) |>
  filter(userId %in% c(13:20)) |>
  select(userId, title, rating) |>
  mutate(title = str_remove(title, ", The"),
         title = str_remove(title, ":.*")) |>
  pivot_wider(names_from = "title", values_from = "rating")

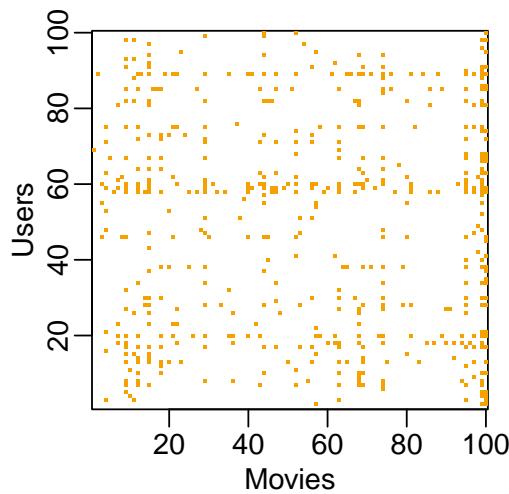
print(tab)

## # A tibble: 5 x 5
##   userId `Pulp Fiction (1994)` `Jurassic Park (1993)` `Silence of the Lambs (1991)` `Forrest Gump (1994)`
##   <int>          <dbl>              <dbl>                  <dbl>              <dbl>
## 1     13            4                NA                  NA             NA
## 2     16            NA               3                  NA             NA
## 3     17            NA               NA                  NA             5
## 4     18            5                3                  NA             5
## 5     19            NA               1                  NA             NA
```

The following plot of the matrix for a random sample of 100 movies and 100 users with yellow indicating a user/movie combination for which we have a rating shows how *sparse* the matrix is:

```
users <- sample(unique(edx$userId), 100)

rafalib::mpar()
edx |>
  filter(userId %in% users) |>
  select(userId, movieId, rating) |>
  mutate(rating = 1) |>
  pivot_wider(names_from = movieId, values_from = rating) |>
  (\(mat) mat[, sample(ncol(mat), 100)]()) |>
  as.matrix() |>
  t() |>
  image(1:100, 1:100, z = _ , xlab = "Movies", ylab = "Users")
```

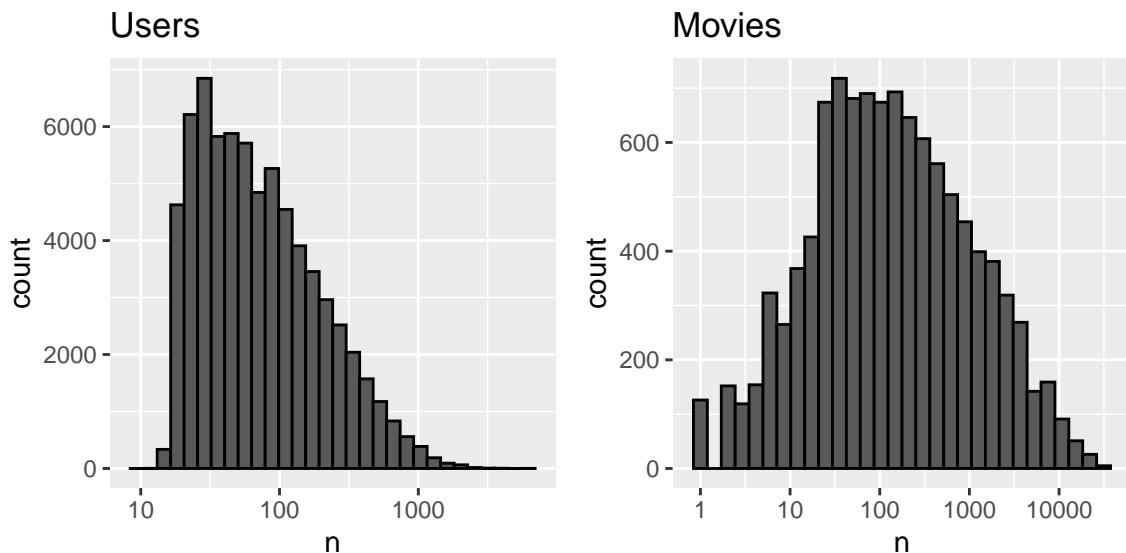


Further observations highlighted there that, as we can see from the distributions the author presented, some movies get rated more than others, and some users are more active than others in rating movies:

```
p1 <- edx |>
  count(movieId) |>
  ggplot(aes(n)) +
  geom_histogram(bins = 30, color = "black") +
  scale_x_log10() +
  ggtitle("Movies")

p2 <- edx |>
  count(userId) |>
  ggplot(aes(n)) +
  geom_histogram(bins = 30, color = "black") +
  scale_x_log10() +
  ggtitle("Users")

gridExtra::grid.arrange(p2, p1, ncol = 2)
```



Finally, we can see that no movies have a rating of 0. Movies are rated from 0.5 to 5.0 in 0.5 increments:

```
#library(dplyr)
s <- edx |> group_by(rating) |>
  summarise(n = n())
print(s)

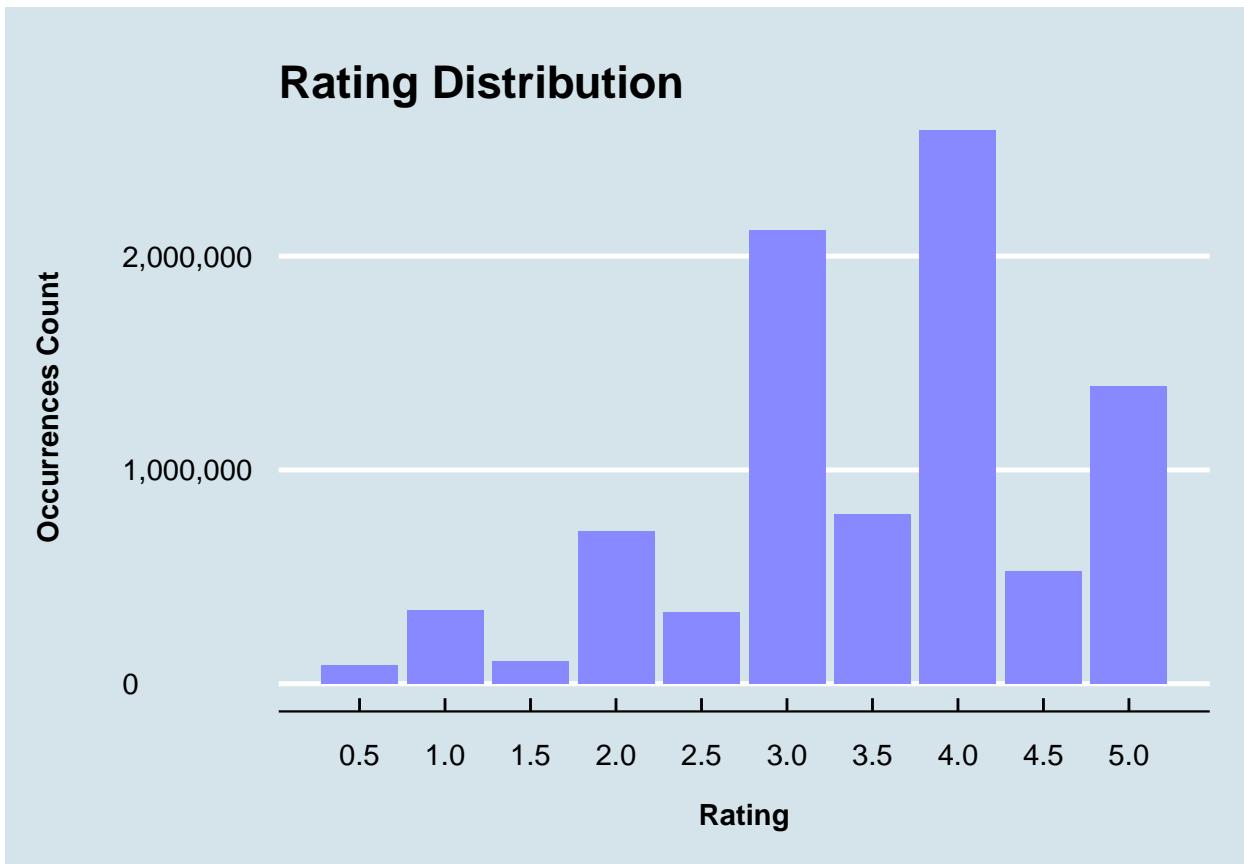
## # A tibble: 10 x 2
##       rating     n
##   <dbl>   <int>
## 1     0.5  85374
## 2     1    345679
## 3     1.5 106426
## 4     2    711422
## 5     2.5 333010
## 6     3   2121240
## 7     3.5 791624
## 8     4   2588430
## 9     4.5 526736
## 10    5   1390114
```

Further analysis of the `edx` dataset have been also inspired by the article mentioned above[6], from which the code and explanatory notes below were cited.

1.1.1.1 Rating distribution plot[6]

The code below demonstrates another way of visualizing the rating distribution:

```
edx |>
  group_by(rating) |>
  summarize(count = n()) |>
  ggplot(aes(x = rating, y = count)) +
  geom_bar(stat = "identity", fill = "#8888ff") +
  ggtitle("Rating Distribution") +
  xlab("Rating") +
  ylab("Occurrences Count") +
  scale_y_continuous(labels = comma) +
  scale_x_continuous(n.breaks = 10) +
  theme_economist() +
  theme(axis.title.x = element_text(vjust = -5, face = "bold"),
        axis.title.y = element_text(vjust = 10, face = "bold"),
        plot.margin = margin(0.7, 0.5, 1, 1.2, "cm"))
```



This graph is another confirmation of what we found out above: rounded ratings occur more often than half-stared ones. The upward trend previously discussed is now perfectly clear, although it seems to top right between the 3 and 4-star ratings lowering the occurrences count afterward. That might be due to users being more hesitant to rate with the highest mark for whichever reasons they might hold[6].

1.1.1.2 Ratings per movie

Movie popularity count[6]

```
print(edx |>
  group_by(movieId) |>
  summarize(count = n()) |>
  slice_head(n = 10)
)

## # A tibble: 10 x 2
##       movieId   count
##   <int>     <int>
## 1      1    23790
## 2      2    10779
## 3      3     7028
## 4      4     1577
## 5      5     6400
## 6      6    12346
## 7      7     7259
## 8      8     821
## 9      9    2278
## 10     10   15187

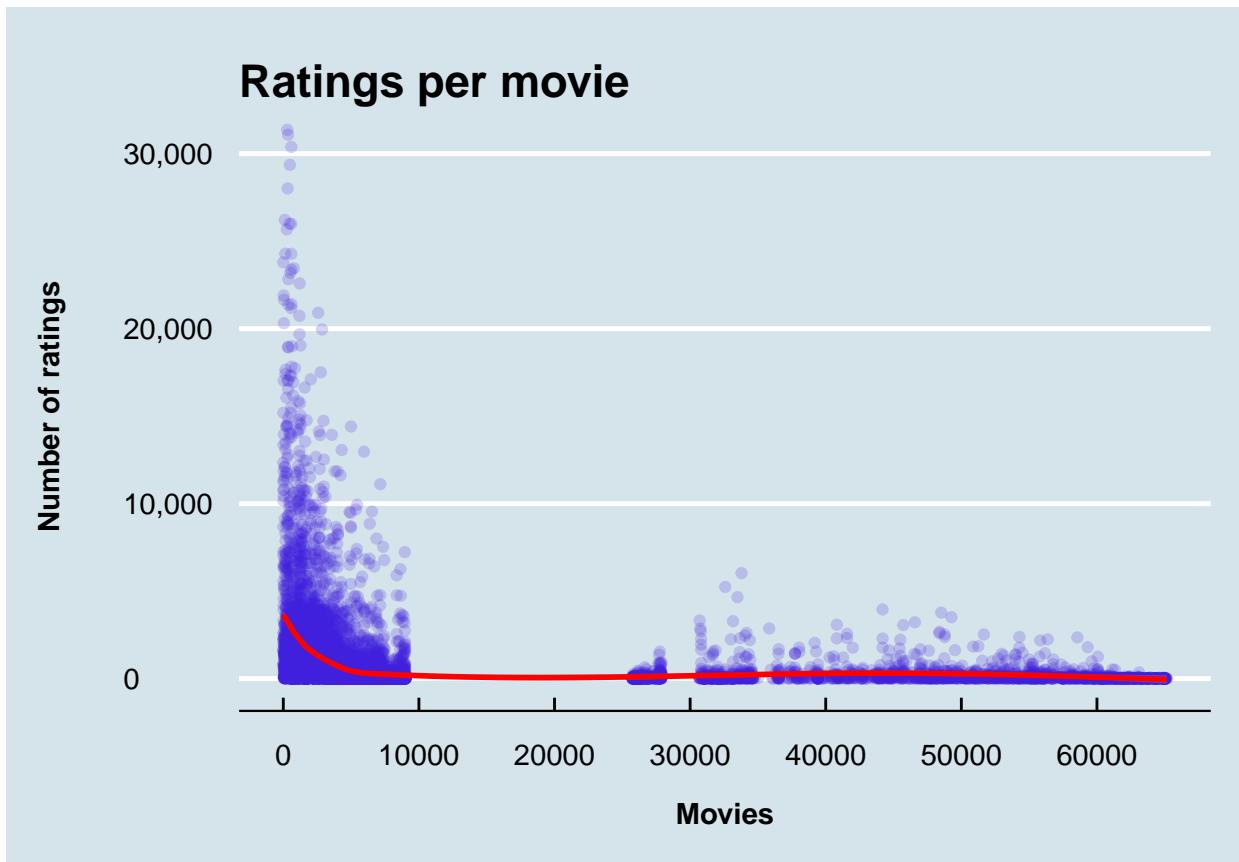
summary(edx |> group_by(movieId) |> summarize(count = n()) |> select(count))

##       count
##   Min.   : 1.0
## 1st Qu.: 30.0
## Median : 122.0
## Mean   : 842.9
## 3rd Qu.: 565.0
## Max.   :31362.0
```

1.1.1.2.1 Ratings per movie plot[6]

```
edx |>
  group_by(movieId) |>
  summarize(count = n()) |>
  ggplot(aes(x = movieId, y = count)) +
  geom_point(alpha = 0.2, color = "#4020dd") +
  geom_smooth(color = "red") +
  ggtitle("Ratings per movie") +
  xlab("Movies") +
  ylab("Number of ratings") +
  scale_y_continuous(labels = comma) +
  scale_x_continuous(n.breaks = 10) +
  theme_economist() +
  theme(axis.title.x = element_text(vjust = -5, face = "bold"),
        axis.title.y = element_text(vjust = 10, face = "bold"),
        plot.margin = margin(0.7, 0.5, 1, 1.2, "cm"))
```

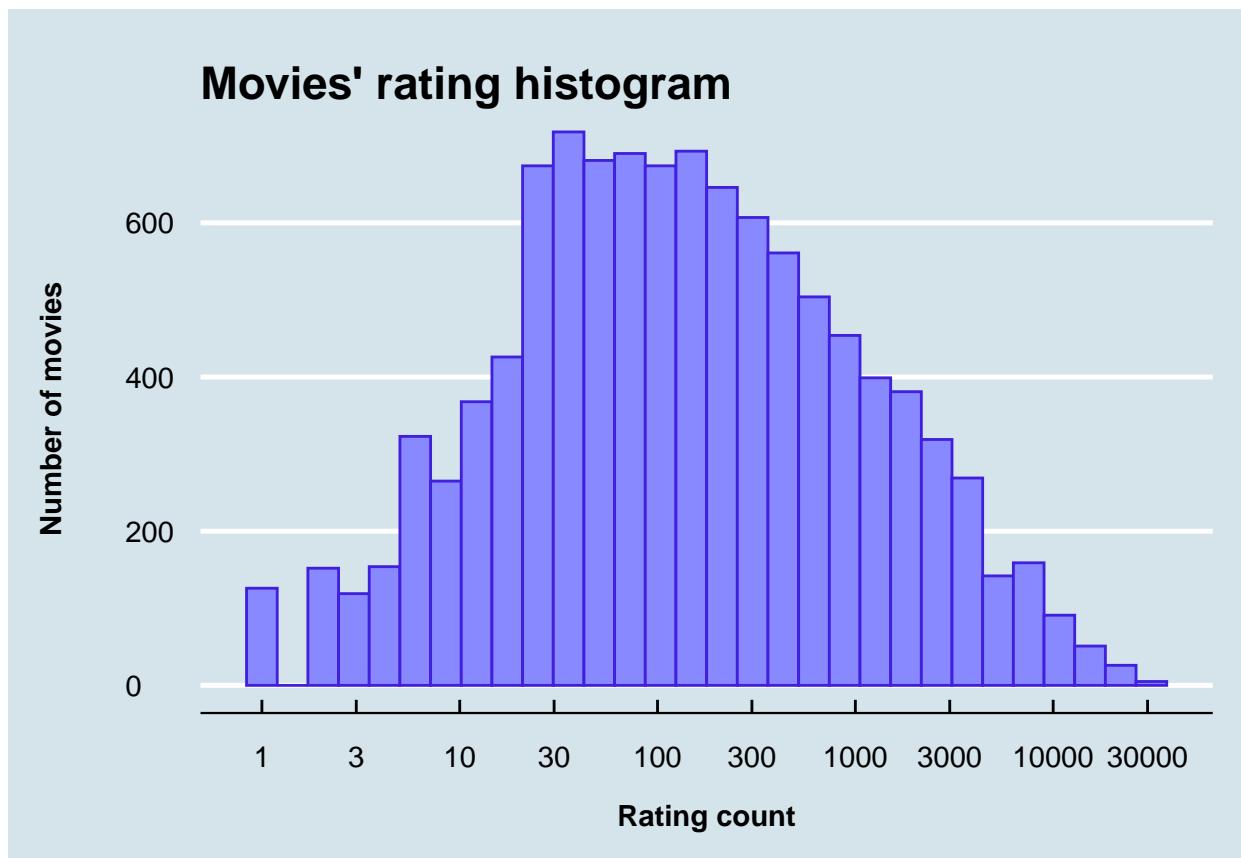
```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```



1.1.1.2.2 Movies' rating histogram[6]

```
edx |>
  group_by(movieId) |>
  summarize(count = n()) |>
  ggplot(aes(x = count)) +
  geom_histogram(fill = "#8888ff", color = "#4020dd") +
  ggtitle("Movies' rating histogram") +
  xlab("Rating count") +
  ylab("Number of movies") +
  scale_y_continuous(labels = comma) +
  scale_x_log10(n.breaks = 10) +
  theme_economist() +
  theme(axis.title.x = element_text(vjust = -5, face = "bold"),
        axis.title.y = element_text(vjust = 10, face = "bold"),
        plot.margin = margin(0.7, 0.5, 1, 1.2, "cm"))
```

‘stat_bin()’ using ‘bins = 30’. Pick better value with ‘binwidth’.



1.1.1.3 Ratings per user[6]

1.1.1.3.1 User rating count (activity measure)

```
print(edx |>
  group_by(userId) |>
  summarize(count = n()) |>
  slice_head(n = 10)
)

## # A tibble: 10 x 2
##   userId count
##     <int> <int>
##   1       1    19
##   2       2    17
##   3       3    31
##   4       4    35
##   5       5    74
##   6       6    39
##   7       7    96
##   8       8   727
##   9       9    21
## 10      10   112
```

1.1.1.3.2 User rating summary

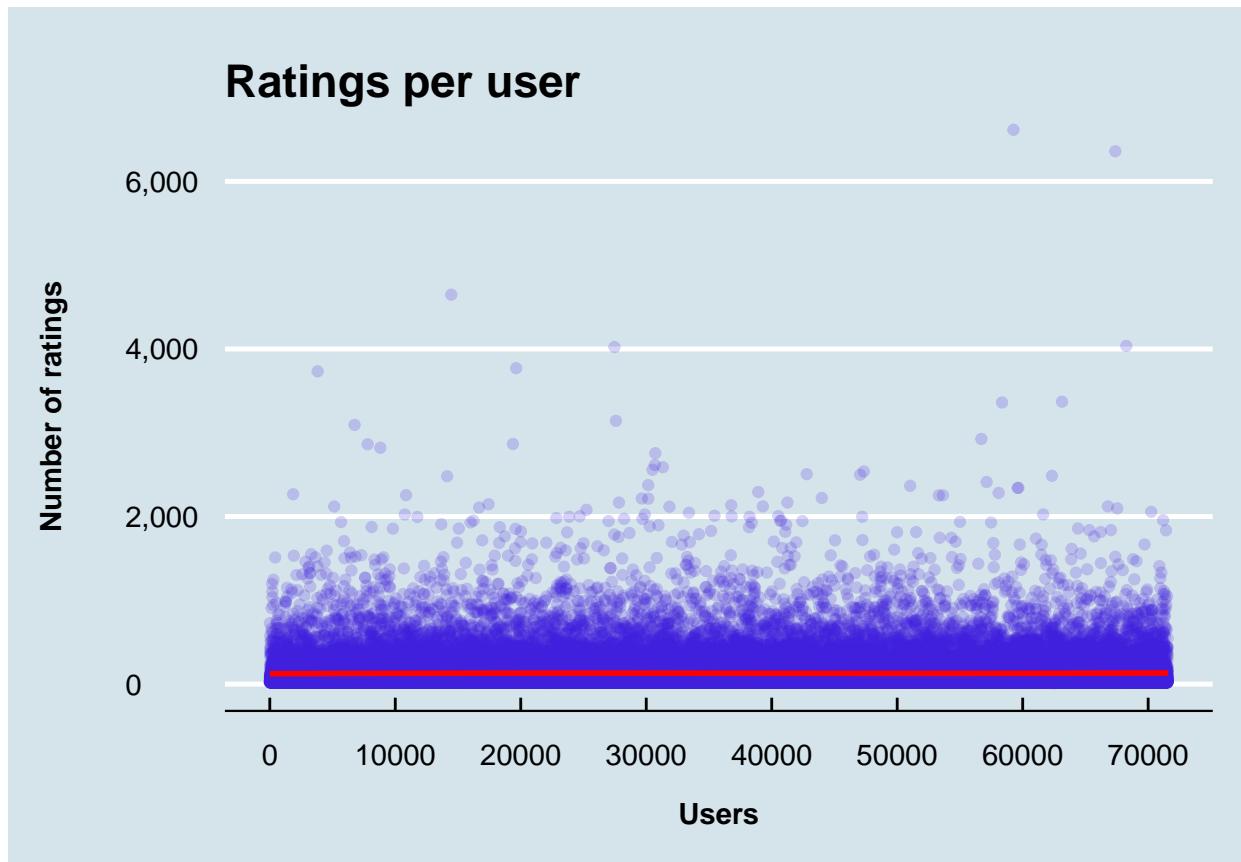
```
summary(edx |> group_by(userId) |> summarize(count = n()) |> select(count))
```

```
##      count
## Min.   : 10.0
## 1st Qu.: 32.0
## Median : 62.0
## Mean   : 128.8
## 3rd Qu.: 141.0
## Max.   :6616.0
```

1.1.1.3.3 Ratings per user plot

```
edx |>
  group_by(userId) |>
  summarize(count = n()) |>
  ggplot(aes(x = userId, y = count)) +
  geom_point(alpha = 0.2, color = "#4020dd") +
  geom_smooth(color = "red") +
  ggtitle("Ratings per user") +
  xlab("Users") +
  ylab("Number of ratings") +
  scale_y_continuous(labels = comma) +
  scale_x_continuous(n.breaks = 10) +
  theme_economist() +
  theme(axis.title.x = element_text(vjust = -5, face = "bold"),
        axis.title.y = element_text(vjust = 10, face = "bold"),
        plot.margin = margin(0.7, 0.5, 1, 1.2, "cm"))
```

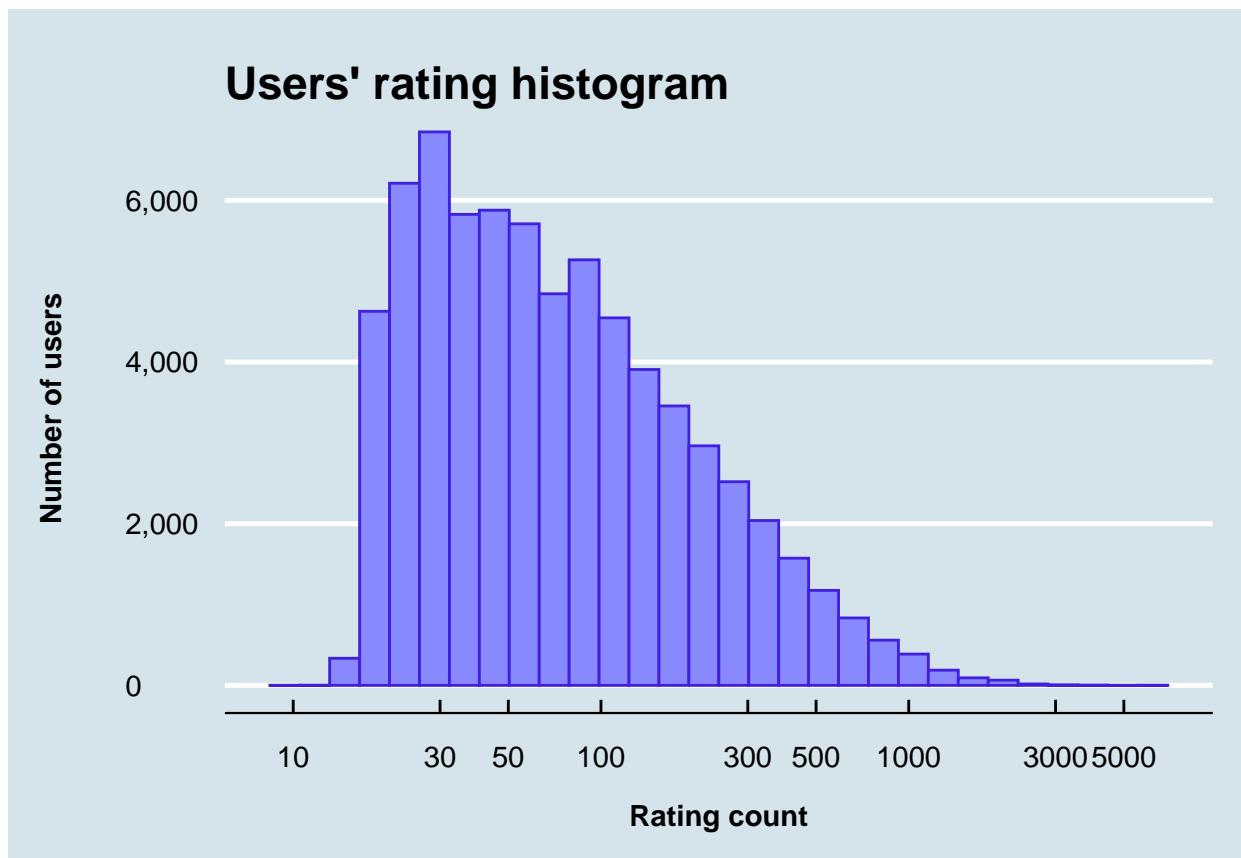
```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```



1.1.1.3.4 Users' rating histogram

```
edx |>
  group_by(userId) |>
  summarize(count = n()) |>
  ggplot(aes(x = count)) +
  geom_histogram(fill = "#8888ff", color = "#4020dd") +
  ggtitle("Users' rating histogram") +
  xlab("Rating count") +
  ylab("Number of users") +
  scale_y_continuous(labels = comma) +
  scale_x_log10(n.breaks = 10) +
  theme_economist() +
  theme(axis.title.x = element_text(vjust = -5, face = "bold"),
        axis.title.y = element_text(vjust = 10, face = "bold"),
        plot.margin = margin(0.7, 0.5, 1, 1.2, "cm"))
```

‘stat_bin()’ using ‘bins = 30’. Pick better value with ‘binwidth’.



2 Methods / Analysis



All the source code of the R-scripts is available on the project's [GitHub repository](#)[7].

2.1 Defining Logging and Time Measuring Helper Functions

First, let's define some helper functions for logging and time-measuring features that we will use in our R scripts. Some of them are listed below:

```
# Logging Helper functions -----
open_logfile <- function(file_name){
  log_file_name <- as.character(Sys.time()) |>
    str_replace_all(':', '_') |>
    str_replace(' ', 'T') |>
    str_c(file_name)

  log_open(file_name = log_file_name)
}

print_start_date <- function(){
  print(date())
  Sys.time()
}

put_start_date <- function(){
  put(date())
  Sys.time()
}

print_end_date <- function(start){
  print(date())
  print(Sys.time() - start)
}

put_end_date <- function(start){
  put(date())
  put(Sys.time() - start)
}

msg.set_arg <- function(msg_template, arg, arg.name = "%1") {
  msg_template |>
    str_replace_all(arg.name, as.character(arg))
}

msg.glue <- function(msg_template, arg, arg.name = "%1"){
  msg_template |>
    msg.set_arg(arg, arg.name) |>
    str_glue()
}

print_log <- function(msg){
  print(str_glue(msg))
}
```

```

put_log <- function(msg){
  put(str_glue(msg))
}

get_log1 <- function(msg_template, arg1) {
  str_glue(str_replace_all(msg_template, "%1", as.character(arg1)))
}
print_log1 <- function(msg_template, arg1){
  print(get_log1(msg_template, arg1))
}
put_log1 <- function(msg_template, arg1){
  put(get_log1(msg_template, arg1))
}

get_log2 <- function(msg_template, arg1, arg2) {
  msg_template |>
    str_replace_all("%1", as.character(arg1)) |>
    str_replace_all("%2", as.character(arg2)) |>
    str_glue()
}
print_log2 <- function(msg_template, arg1, arg2){
  print(get_log1(msg_template, arg1, arg2))
}
put_log2 <- function(msg_template, arg1, arg2){
  put(get_log1(msg_template, arg1, arg2))
}

# ...

```



The full source code of these functions is available in the [Logging Helper functions](#) section of the [capstone-movielens.main.R](#) script on *GitHub*.

2.2 Preparing train and test datasets

We will split the `edx` dataset into a training set, which we will use to build and train our models, and a test set in which we will compute the accuracy of our predictions, the way described in [Section 23.1.1 MovieLens data](#) of the *Course Textbook* mentioned above[5]. We will also use the *5-Fold Cross Validation* method as described in [Section 29.6 Cross validation](#) of the *Course Textbook*. To prepare datasets for processing, we will use the following functions, specifically designed for these operations:

- `make_source_datasets`
- `init_source_datasets`



The full source code of the function listed above is available in the [Initialize input datasets](#) section of the `data.helper.functions.R` script on *Github*.

2.2.1 The `make_source_datasets` function

Let's take a closer look at the objects we will receive as a result of executing this function.

```
make_source_datasets <- function(){  
  # ...  
  list(edx_CV = edx_CV,  
       edx.mx = edx.mx,  
       edx.sgr = edx.sgr,  
       tuning_sets = tuning_sets,  
       movie_map = movie_map,  
       date_days_map = date_days_map)  
}
```

2.2.1.1 `edx.mx` Matrix Object

We will use the array representation described in [Section 17.5 of the Textbook](#), for the training data: we denote ranking for movie j by user i as $y_{i,j}$. To create this matrix, we use `tidyverse::pivot_wider` function:

```
put_log("Function: `make_source_datasets`: Creating Rating Matrix from `edx` dataset...")  
edx.mx <- edx |>  
  mutate(userId = factor(userId),  
         movieId = factor(movieId)) |>  
  select(movieId, userId, rating) |>  
  pivot_wider(names_from = movieId, values_from = rating) |>  
  column_to_rownames("userId") |>  
  as.matrix()  
  
put_log("Function: `make_source_datasets`:  
Matrix created: `edx.mx` of the following dimensions:")
```

```

str(edx.mx)

##  num [1:69878, 1:10677] 5 NA NA NA NA NA NA NA NA NA ...
## - attr(*, "dimnames")=List of 2
##   ..$ : chr [1:69878] "1" "2" "3" "4" ...
##   ..$ : chr [1:10677] "122" "185" "292" "316" ...

```

2.2.1.2 edx.sgr Object

To account for the Movie Genre Effect more accurately, we need a dataset with split rows for movies belonging to multiple genres:

```

put_log("Function: `make_source_datasets`:
To account for the Movie Genre Effect, we need a dataset with split rows
for movies belonging to multiple genres.")
edx.sgr <- splitGenreRows(edx)

```

```
str(edx.sgr)
```

```

## #tibble [23,371,423 x 6] (S3:tbl_df/tbl/data.frame)
## $ userId    : int [1:23371423] 1 1 1 1 1 1 1 1 1 1 ...
## $ movieId   : int [1:23371423] 122 122 185 185 185 292 292 292 292 316 ...
## $ rating    : num [1:23371423] 5 5 5 5 5 5 5 5 5 5 ...
## $ timestamp: int [1:23371423] 838985046 838985046 838983525 838983525 838983525 838983421 838983421
## $ title     : chr [1:23371423] "Boomerang (1992)" "Boomerang (1992)" "Net, The (1995)" "Net, The (1995" ...
## $ genres    : chr [1:23371423] "Comedy" "Romance" "Action" "Crime" ...

```

```
summary(edx.sgr)
```

	userId	movieId	rating	timestamp	title	genres
## Min.	: 1	Min. : 1	Min. : 0.500	Min. : 7.897e+08	Length:23371423	Length:23371423
## 1st Qu.:	18140	1st Qu.: 616	1st Qu.: 3.000	1st Qu.: 9.472e+08	Class :character	Class :character
## Median :	35784	Median : 1748	Median : 4.000	Median : 1.042e+09	Mode :character	Mode :character
## Mean :	35886	Mean : 4277	Mean : 3.527	Mean : 1.035e+09		
## 3rd Qu.:	53638	3rd Qu.: 3635	3rd Qu.: 4.000	3rd Qu.: 1.131e+09		
## Max. :	71567	Max. : 65133	Max. : 5.000	Max. : 1.231e+09		

Note that we use the `splitGenreRows` function to split rows of the original dataset:

```

splitGenreRows <- function(data){
  put("Splitting dataset rows related to multiple genres...")
  start <- put_start_date()
  gs_splitted <- data |>
    separate_rows(genres, sep = "\\\\|")
  put("Dataset rows related to multiple genres have been splitted to have single genre per row.")
  put_end_date(start)
  gs_splitted
}

```



The source code of the function mentioned above is also available in the [Initialize input datasets](#) section of the `data.helper.functions.R` script on *GitHub*.

2.2.1.3 movie_map Object

To be able to map movie IDs to titles we create the following lookup table:

```
movie_map <- edx |> select(movieId, title, genres) |>
  distinct(movieId, .keep_all = TRUE)

  put_log("Function: `make_source_datasets`: Dataset created: movie_map")

  str(movie_map)

## 'data.frame': 10677 obs. of 3 variables:
## $ movieId: int 122 185 292 316 329 355 356 362 364 370 ...
## $ title  : chr "Boomerang (1992)" "Net, The (1995)" "Outbreak (1995)" "Stargate (1994)" ...
## $ genres : chr "Comedy|Romance" "Action|Crime|Thriller" "Action|Drama|Sci-Fi|Thriller" "Action|Adv...

summary(movie_map)

##      movieId          title           genres
## Min.   :    1   Length:10677   Length:10677
## 1st Qu.: 2754   Class :character   Class :character
## Median : 5434   Mode  :character   Mode  :character
## Mean   :13105
## 3rd Qu.: 8710
## Max.   :65133
```

Note that titles cannot be considered unique, so we can't use them as IDs[5].

2.2.1.4 date_days_map Object

We have a `timestamp` field in the `edx` dataset. To be able to map the date, year, and number of days since the earliest record in the `edx` dataset with the corresponding value in this field, we create the following lookup table:

```
put_log("Function: `make_source_datasets`: Creating Date-Days Map dataset...")
date_days_map <- edx |>
  mutate(date_time = as_datetime(timestamp)) |>
  mutate(date = as_date(date_time)) |>
  mutate(year = year(date_time)) |>
  mutate(days = as.integer(date - min(date))) |>
  select(timestamp, date_time, date, year, days) |>
  distinct(timestamp, .keep_all = TRUE)

  put_log("Function: `make_source_datasets`: Dataset created: date_days_map")

  str(date_days_map)

## 'data.frame': 6519590 obs. of 5 variables:
## $ timestamp: int 838985046 838983525 838983421 838983392 838984474 838983653 838984885 838983707 838984123 ...
```

```

## $ date_time: POSIXct, format: "1996-08-02 11:24:06" "1996-08-02 10:58:45" "1996-08-02 10:57:01" "1996-08-02 10:57:01" ...
## $ date      : Date, format: "1996-08-02" "1996-08-02" "1996-08-02" "1996-08-02" ...
## $ year      : num  1996 1996 1996 1996 1996 ...
## $ days      : int  571 571 571 571 571 571 571 571 571 571 ...

summary(date_days_map)

##   timestamp        date_time          date       year    days
## Min.    :7.897e+08 Min.   :1995-01-09 11:46:49.00 Min.   :1995  1995  Min.   :
## 1st Qu.:9.783e+08 1st Qu.:2001-01-01 05:05:01.75 1st Qu.:2001 2001  1st Qu.:2001
## Median :1.091e+09 Median :2004-08-03 01:08:18.50 Median :2004 2004  Median :2004
## Mean   :1.066e+09 Mean   :2003-10-10 23:15:02.07 Mean   :2003 2003  Mean   :311
## 3rd Qu.:1.152e+09 3rd Qu.:2006-07-04 20:41:57.50 3rd Qu.:2006 2006  3rd Qu.:2006
## Max.   :1.231e+09 Max.   :2009-01-05 05:02:16.00 Max.   :2009 2009  Max.   :511

```

2.2.1.5 edx_CV Object

Here we have a list of sample objects we need to perform the *5-Fold Cross Validation* as explained in Section 29.6.1 K-fold cross validation of the *Course Textbook*:

```

start <- put_start_date()
edx_CV <- lapply(kfold_index,  function(fold_i){

  put_log1("Method `make_source_datasets`:
Creating K-Fold Cross Validation Datasets, Fold %1", fold_i)

  #> We split the initial datasets into training sets, which we will use to build
  #> and train our models, and validation sets in which we will compute the accuracy
  #> of our predictions, the way described in the `Section 23.1.1 MovieLens data`-
  #> (https://rafaelab.dfci.harvard.edu/dsbook-part-2/highdim/regularization.html#movielens-data)
  #> of the Course Textbook.

  split_sets <- edx |>
    sample_train_validation_sets(fold_i*1000)

  train_set <- split_sets$train_set
  validation_set <- split_sets$validation_set

  put_log("Function: `make_source_datasets`:
Sampling 20% from the split-row version of the `edx` dataset...")
  split_sets.gs <- edx.sgr |>
    sample_train_validation_sets(fold_i*2000)

  train.sgr <- split_sets.gs$train_set
  validation.sgr <- split_sets.gs$validation_set

  # put_log("Function: `make_source_datasets`: Dataset created: validation.sgr")
  # put(summary(validation.sgr))

  #> We will use the array representation described in `Section 17.5 of the Textbook`-
  #> (https://rafaelab.dfci.harvard.edu/dsbook-part-2/linear-models/treatment-effect-models.html#sec-a)
  #> for the training data.
}

```

```

#> To create this matrix, we use `tidyverse::pivot_wider` function:

put_log("Function: `make_source_datasets`: Creating Rating Matrix from Train Set...")
train_mx <- train_set |>
  mutate(userId = factor(userId),
         movieId = factor(movieId)) |>
  select(movieId, userId, rating) |>
  pivot_wider(names_from = movieId, values_from = rating) |>
  column_to_rownames("userId") |>
  as.matrix()

put_log("Function: `make_source_datasets`:
Matrix created: `train_mx` of the following dimensions:")
put(dim(train_mx))

list(train_set = train_set,
     train_mx = train_mx,
     train.sgr = train.sgr,
     validation_set = validation_set)
})

put_end_date(start)
put_log("Function: `make_source_datasets`:
Set of K-Fold Cross Validation datasets created: edx_CV")

```

```
str(edx_CV)
```

```

## List of 5
## $ :List of 4
##   ..$ train_set      :'data.frame':  7172311 obs. of  6 variables:
##     ...$ userId      : int [1:7172311] 1 1 1 1 1 1 1 1 1 ...
##     ...$ movieId     : int [1:7172311] 122 185 292 329 356 362 364 370 420 466 ...
##     ...$ rating      : num [1:7172311] 5 5 5 5 5 5 5 5 5 ...
##     ...$ timestamp   : int [1:7172311] 838985046 838983525 838983421 838983392 838983653 838984885 838984885 ...
##     ...$ title       : chr [1:7172311] "Boomerang (1992)" "Net, The (1995)" "Outbreak (1995)" "Star Trek: Generations (1994)" ...
##     ...$ genres      : chr [1:7172311] "Comedy|Romance" "Action|Crime|Thriller" "Action|Drama|Sci-Fi|Thriller" ...
##   ..$ train_mx      : num [1:69878, 1:10677] 5 NA NA NA NA NA NA NA NA ...
##   ...- attr(*, "dimnames")=List of 2
##     ...$ : chr [1:69878] "1" "2" "3" "4" ...
##     ...$ : chr [1:10677] "122" "185" "292" "329" ...
##   ..$ train.sgr     : tibble [18,669,190 x 6] (S3: tbl_df/tbl/data.frame)
##     ...$ userId      : int [1:18669190] 1 1 1 1 1 1 1 1 1 ...
##     ...$ movieId     : int [1:18669190] 122 122 185 185 292 292 292 292 316 316 ...
##     ...$ rating      : num [1:18669190] 5 5 5 5 5 5 5 5 ...
##     ...$ timestamp   : int [1:18669190] 838985046 838985046 838983525 838983525 838983421 838983421 838983421 ...
##     ...$ title       : chr [1:18669190] "Boomerang (1992)" "Boomerang (1992)" "Net, The (1995)" "Net, The (1995)" ...
##     ...$ genres      : chr [1:18669190] "Comedy" "Romance" "Action" "Crime" ...
##   ..$ validation_set:'data.frame':  1827744 obs. of  6 variables:
##     ...$ userId      : int [1:1827744] 1 1 1 1 2 2 2 2 3 3 ...
##     ...$ movieId     : int [1:1827744] 316 355 377 588 260 376 648 1049 110 1252 ...
##     ...$ rating      : num [1:1827744] 5 5 5 5 3 2 3 4.5 4 ...
##     ...$ timestamp   : int [1:1827744] 838983392 838984474 838983834 838983339 868244562 868245920 868245920 ...
##     ...$ title       : chr [1:1827744] "Stargate (1994)" "Flintstones, The (1994)" "Speed (1994)" "Aladdin (1992)" ...
##     ...$ genres      : chr [1:1827744] "Action|Adventure|Sci-Fi" "Children|Comedy|Fantasy" "Action|Romantic" ...

```

```

## $ :List of 4
## ..$ train_set      :'data.frame': 7172306 obs. of 6 variables:
## ...$ userId       : int [1:7172306] 1 1 1 1 1 1 1 1 1 ...
## ...$ movieId      : int [1:7172306] 122 185 292 316 329 355 356 364 370 377 ...
## ...$ rating       : num [1:7172306] 5 5 5 5 5 5 5 5 5 5 ...
## ...$ timestamp    : int [1:7172306] 838985046 838983525 838983421 838983392 838983392 838984474 838984474 838984474 838984474 838984474 ...
## ...$ title        : chr [1:7172306] "Boomerang (1992)" "Net, The (1995)" "Outbreak (1995)" "Stargate ...
## ...$ genres       : chr [1:7172306] "Comedy|Romance" "Action|Crime|Thriller" "Action|Drama|Sci-Fi|Th ...
## ..$ train_mx     : num [1:69878, 1:10677] 5 NA NA NA NA NA NA NA NA NA ...
## ...- attr(*, "dimnames")=List of 2
## ... .$. : chr [1:69878] "1" "2" "3" "4" ...
## ... .$. : chr [1:10677] "122" "185" "292" "316" ...
## ..$ train.sgr    : tibble [18,669,201 x 6] (S3: tbl_df/tbl/data.frame)
## ...$ userId       : int [1:18669201] 1 1 1 1 1 1 1 1 1 ...
## ...$ movieId      : int [1:18669201] 122 122 185 185 185 292 292 316 316 329 ...
## ...$ rating       : num [1:18669201] 5 5 5 5 5 5 5 5 5 5 ...
## ...$ timestamp    : int [1:18669201] 838985046 838985046 838983525 838983525 838983525 838983525 838983421 838983421 838983421 838983421 ...
## ...$ title        : chr [1:18669201] "Boomerang (1992)" "Boomerang (1992)" "Net, The (1995)" "Net, Th ...
## ...$ genres       : chr [1:18669201] "Comedy" "Romance" "Action" "Crime" ...
## ..$ validation_set:'data.frame': 1827749 obs. of 6 variables:
## ...$ userId       : int [1:1827749] 1 1 1 1 2 2 2 2 3 3 ...
## ...$ movieId      : int [1:1827749] 362 520 539 594 539 590 733 1210 1252 1408 ...
## ...$ rating       : num [1:1827749] 5 5 5 5 3 5 3 4 4 3.5 ...
## ...$ timestamp    : int [1:1827749] 838984885 838984679 838984068 838984679 868246262 868245608 868245608 868245608 868245608 ...
## ...$ title        : chr [1:1827749] "Jungle Book, The (1994)" "Robin Hood: Men in Tights (1993)" "Sl ...
## ...$ genres       : chr [1:1827749] "Adventure|Children|Romance" "Comedy" "Comedy|Drama|Romance" "An ...
## $ :List of 4
## ..$ train_set      :'data.frame': 7172307 obs. of 6 variables:
## ...$ userId       : int [1:7172307] 1 1 1 1 1 1 1 1 1 ...
## ...$ movieId      : int [1:7172307] 122 185 292 316 329 355 362 370 377 420 ...
## ...$ rating       : num [1:7172307] 5 5 5 5 5 5 5 5 5 5 ...
## ...$ timestamp    : int [1:7172307] 838985046 838983525 838983421 838983392 838983392 838984474 838984474 838984474 838984474 838984474 ...
## ...$ title        : chr [1:7172307] "Boomerang (1992)" "Net, The (1995)" "Outbreak (1995)" "Stargate ...
## ...$ genres       : chr [1:7172307] "Comedy|Romance" "Action|Crime|Thriller" "Action|Drama|Sci-Fi|Th ...
## ..$ train_mx     : num [1:69878, 1:10677] 5 NA NA NA NA NA NA NA NA ...
## ...- attr(*, "dimnames")=List of 2
## ... .$. : chr [1:69878] "1" "2" "3" "4" ...
## ... .$. : chr [1:10677] "122" "185" "292" "316" ...
## ..$ train.sgr    : tibble [18,669,195 x 6] (S3: tbl_df/tbl/data.frame)
## ...$ userId       : int [1:18669195] 1 1 1 1 1 1 1 1 1 ...
## ...$ movieId      : int [1:18669195] 122 122 185 185 185 292 292 292 316 329 ...
## ...$ rating       : num [1:18669195] 5 5 5 5 5 5 5 5 5 5 ...
## ...$ timestamp    : int [1:18669195] 838985046 838985046 838983525 838983525 838983525 838983525 838983421 838983421 838983421 838983421 ...
## ...$ title        : chr [1:18669195] "Boomerang (1992)" "Boomerang (1992)" "Net, The (1995)" "Net, Th ...
## ...$ genres       : chr [1:18669195] "Comedy" "Romance" "Action" "Crime" ...
## ..$ validation_set:'data.frame': 1827748 obs. of 6 variables:
## ...$ userId       : int [1:1827748] 1 1 1 1 2 2 2 2 3 3 ...
## ...$ movieId      : int [1:1827748] 356 364 539 616 590 719 780 786 151 213 ...
## ...$ rating       : num [1:1827748] 5 5 5 5 3 3 3 4.5 5 ...
## ...$ timestamp    : int [1:1827748] 838983653 838983707 838984068 838984941 868245608 868246191 868246191 868246191 868246191 ...
## ...$ title        : chr [1:1827748] "Forrest Gump (1994)" "Lion King, The (1994)" "Sleepless in Seat ...
## ...$ genres       : chr [1:1827748] "Comedy|Drama|Romance|War" "Adventure|Animation|Children|Drama|M ...
## $ :List of 4
## ..$ train_set      :'data.frame': 7172311 obs. of 6 variables:

```

```

## ... .$.userId : int [1:7172311] 1 1 1 1 1 1 1 1 1 1 ...
## ... .$.movieId : int [1:7172311] 122 185 292 316 329 355 356 362 364 370 ...
## ... .$.rating : num [1:7172311] 5 5 5 5 5 5 5 5 5 5 ...
## ... .$.timestamp: int [1:7172311] 838985046 838983525 838983421 838983392 838983392 838984474 838983653 838983653 838983653 838983653 ...
## ... .$.title : chr [1:7172311] "Boomerang (1992)" "Net, The (1995)" "Outbreak (1995)" "Stargate (1994)" "The Net (1995)" "The Outbreak (1995)" "The Stargate (1994)" ...
## ... .$.genres : chr [1:7172311] "Comedy|Romance" "Action|Crime|Thriller" "Action|Drama|Sci-Fi|Thriller" "Action|Drama|Sci-Fi|Thriller" "Action|Drama|Sci-Fi|Thriller" "Action|Drama|Sci-Fi|Thriller" "Action|Drama|Sci-Fi|Thriller" "Action|Drama|Sci-Fi|Thriller" "Action|Drama|Sci-Fi|Thriller" "Action|Drama|Sci-Fi|Thriller" ...
## ... $.train_mx : num [1:69878, 1:10677] 5 NA NA NA NA NA NA NA NA NA ...
## ... - attr(*, "dimnames")=List of 2
## ... ... $. : chr [1:69878] "1" "2" "3" "4" ...
## ... ... $. : chr [1:10677] "122" "185" "292" "316" ...
## ... $.train.sgr : tibble [18,669,192 x 6] (S3: tbl_df/tbl/data.frame)
## ... ... $.userId : int [1:18669192] 1 1 1 1 1 1 1 1 1 ...
## ... ... $.movieId : int [1:18669192] 122 122 185 185 292 292 316 316 329 ...
## ... ... $.rating : num [1:18669192] 5 5 5 5 5 5 5 5 5 ...
## ... ... $.timestamp: int [1:18669192] 838985046 838985046 838983525 838983525 838983421 838983421 838983653 838983653 838983653 ...
## ... ... $.title : chr [1:18669192] "Boomerang (1992)" "Boomerang (1992)" "Net, The (1995)" "Net, The (1995)" "Outbreak (1995)" "Stargate (1994)" ...
## ... ... $.genres : chr [1:18669192] "Comedy" "Romance" "Action" "Thriller" ...
## ... $.validation_set:'data.frame': 1827744 obs. of 6 variables:
## ... ... $.userId : int [1:1827744] 1 1 1 1 2 2 2 2 3 3 ...
## ... ... $.movieId : int [1:1827744] 377 520 588 616 110 648 1049 1356 1148 1276 ...
## ... ... $.rating : num [1:1827744] 5 5 5 5 5 2 3 3 4 3.5 ...
## ... ... $.timestamp: int [1:1827744] 838983834 838984679 838983339 838984941 868245777 868244699 868245777 868244699 868245777 ...
## ... ... $.title : chr [1:1827744] "Speed (1994)" "Robin Hood: Men in Tights (1993)" "Aladdin (1992)" "The Lion King (1994)" ...
## ... ... $.genres : chr [1:1827744] "Action|Romance|Thriller" "Comedy" "Adventure|Animation|Children" "Romantic|Thriller" ...
## ... $ :List of 4
## ... ... $.train_set : 'data.frame': 7172301 obs. of 6 variables:
## ... ... ... $.userId : int [1:7172301] 1 1 1 1 1 1 1 1 1 ...
## ... ... ... $.movieId : int [1:7172301] 122 185 292 316 355 356 364 370 420 466 ...
## ... ... ... $.rating : num [1:7172301] 5 5 5 5 5 5 5 5 5 ...
## ... ... ... $.timestamp: int [1:7172301] 838985046 838983525 838983421 838983392 838984474 838983653 838983653 838983653 838983653 ...
## ... ... ... $.title : chr [1:7172301] "Boomerang (1992)" "Net, The (1995)" "Outbreak (1995)" "Stargate (1994)" ...
## ... ... ... $.genres : chr [1:7172301] "Comedy|Romance" "Action|Crime|Thriller" "Action|Drama|Sci-Fi|Thriller" "Action|Drama|Sci-Fi|Thriller" ...
## ... ... $.train_mx : num [1:69878, 1:10677] 5 NA NA NA NA NA NA NA NA ...
## ... ... - attr(*, "dimnames")=List of 2
## ... ... ... $. : chr [1:69878] "1" "2" "3" "4" ...
## ... ... ... $. : chr [1:10677] "122" "185" "292" "316" ...
## ... ... $.train.sgr : tibble [18,669,194 x 6] (S3: tbl_df/tbl/data.frame)
## ... ... ... $.userId : int [1:18669194] 1 1 1 1 1 1 1 1 1 ...
## ... ... ... $.movieId : int [1:18669194] 122 122 185 185 292 292 316 329 329 355 ...
## ... ... ... $.rating : num [1:18669194] 5 5 5 5 5 5 5 5 5 ...
## ... ... ... $.timestamp: int [1:18669194] 838985046 838985046 838983525 838983525 838983421 838983421 838983653 838983653 838983653 ...
## ... ... ... $.title : chr [1:18669194] "Boomerang (1992)" "Boomerang (1992)" "Net, The (1995)" "Net, The (1995)" "Outbreak (1995)" "Stargate (1994)" ...
## ... ... ... $.genres : chr [1:18669194] "Comedy" "Romance" "Crime" "Thriller" ...
## ... ... $.validation_set:'data.frame': 1827754 obs. of 6 variables:
## ... ... ... $.userId : int [1:1827754] 1 1 1 1 2 2 2 2 3 3 ...
## ... ... ... $.movieId : int [1:1827754] 329 362 377 594 110 376 539 736 1252 1408 ...
## ... ... ... $.rating : num [1:1827754] 5 5 5 5 3 3 3 4 3.5 ...
## ... ... ... $.timestamp: int [1:1827754] 838983392 838984885 838983834 838984679 868245777 868245920 868245777 868245920 868245777 ...
## ... ... ... $.title : chr [1:1827754] "Star Trek: Generations (1994)" "Jungle Book, The (1994)" "Speed (1994)" "The Lion King (1994)" ...
## ... ... ... $.genres : chr [1:1827754] "Action|Adventure|Drama|Sci-Fi" "Adventure|Children|Romance" "Action|Romance" ...

```



This code snippet is a part of the `make_source_datasets` function code described above.

Note that we used the `sample_train_validation_sets` function call to split the original dataset (`edx` in this case):

```
split_sets <- edx |>
  sample_train_validation_sets(fold_i*1000)
```

which returns a pair of train/validation sets:

```
sample_train_validation_sets <- function(data, seed){
  put_log("Function: `sample_train_validation_sets`: Sampling 20% of the `data` data...")
  set.seed(seed)
  validation_ind <-
    sapply(splitByUser(data),
      function(i) sample(i, ceiling(length(i)*.2))) |>
    unlist() |>
    sort()

  put_log("Function: `sample_train_validation_sets`:
Extracting 80% of the original `data` not used for the Validation Set,
excluding data for users who provided no more than a specified number of ratings: {min_nratings}.")

  train_set <- data[-validation_ind,]

  put_log("Function: `sample_train_validation_sets`: Dataset created: train_set")
  put(summary(train_set))

  put_log("Function: `sample_train_validation_sets`:
To make sure we don't include movies in the Training Set that should not be there,
we exclude entries using the semi_join function from the Validation Set.")
  tmp.data <- data[validation_ind,]

  validation_set <- tmp.data |>
    semi_join(train_set, by = "movieId") |>
    semi_join(train_set, by = "userId") |>
    as.data.frame()

  # Add rows excluded from `validation_set` into `train_set`
  tmp.excluded <- anti_join(tmp.data, validation_set)
  train_set <- rbind(train_set, tmp.excluded)

  put_log("Function: `sample_train_validation_sets`: Dataset created: validation_set")
  put(summary(validation_set))

  # CV train & test sets Consistency Test
  validation.left_join.Nas <- train_set |>
    mutate(tst.col = rating) |>
    select(userId, movieId, tst.col) |>
    data.consistency.test(validation_set)

  put_log("Function: `sample_train_validation_sets`:
Below are the data consistency verification results")
  put(validation.left_join.Nas)

  # Return result datasets -----
```

```
    list(train_set = train_set,
         validation_set = validation_set)
}
```



The `sample_train_validation_sets` function is defined in the same script as the `make_source_datasets`⁴ one, from where it is called.

2.2.2 Common Helper Functions

For our further analysis, we are going to use the following *common helper functions*:

2.2.2.1 clamp function

As explained in [Section 24.4 User effects](#) of the *Course Textbook* we know ratings can't be below 0.5 or above 5. For this reason, we will use the `clamp` function described in that section:

```
clamp <- function(x, min = 0.5, max = 5) pmax(pmin(x, max), min)
```

2.2.2.2 Functions to calculate (*Root*) Mean Squared Error

We will need the following functions to calculate (*R*)MSEs:

```
mse <- function(r) mean(r^2)

mse_cv <- function(r_list) {
  mses <- sapply(r_list, mse(r))
  mean(mses)
}

rmse <- function(r) sqrt(mse(r))
# rmse_cv <- function(r_list) sqrt(mse_cv(r_list))

rmse2 <- function(true_ratings, predicted_ratings) {
  rmse(true_ratings - predicted_ratings)
}
```



All the *common helper functions*, including those described above, are defined in the [common-helper.functions.R](#) script on *GitHub*.

2.3 Overall Mean Rating (Naive) Model

Let's begin our analysis by evaluating the simplest model described in [Section 23.3 The First Model of the Course Textbook](#), and then gradually refine it through further research. It is about a model that assumes the same rating for all movies and users with all the differences explained by random variation would look as follows:

$$Y_{i,j} = \mu + \varepsilon_{i,j}$$

with $\varepsilon_{i,j}$ independent errors sampled from the same distribution centered at 0 and μ the *true* rating for all movies.

We know that the estimate that minimizes the RMSE is the least squares estimate of μ and, in this case, is the average of all ratings:

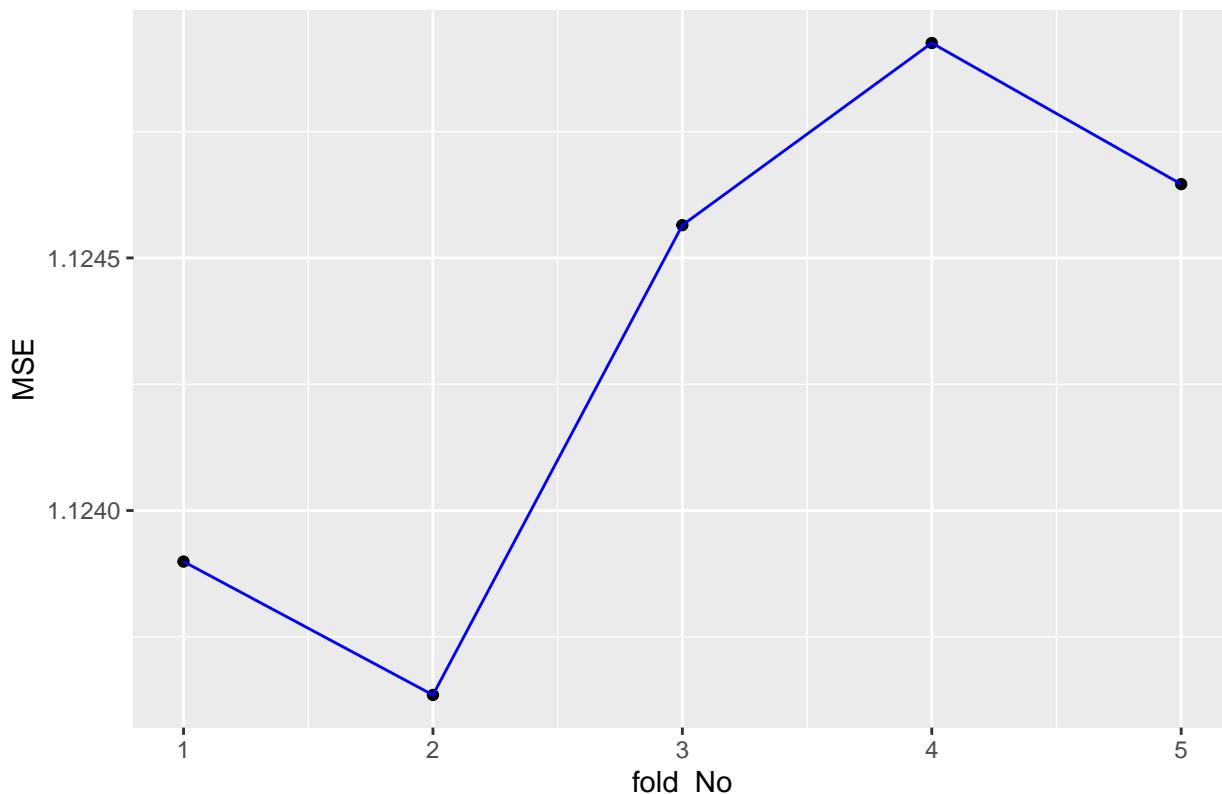
```
mu <- mean(edx$rating)
print(mu)
```

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

If we predict all unknown ratings with $\hat{\mu}$, we obtain the following RMSE:

```
mu.MSEs <- naive_model_MSEs(mu)
data.frame(fold_No = 1:5, MSE = mu.MSEs) |>
  data.plot(title = "MSE results of the 5-fold CV method applied to the Overall Mean Rating Model",
            xname = "fold_No",
            yname = "MSE")
```

MSE results of the 5-fold CV method applied to the Overall Mean Rating



```
mu.RMSE <- sqrt(mean(mu.MSEs))  
mu.RMSE
```

```
## [1] 1.060346
```



For the *Mean Squared Error* data visualization we used `data.plot` function] defined in the [Data Visualization](#) section of the `data.helper.function.R` script.

```
data.plot <- function(data,  
                      title,  
                      xname,  
                      yname,  
                      xlabel = NULL,  
                      ylabel = NULL,  
                      line_col = "blue",  
                      # scale = 1,  
                      normalize = FALSE) {  
  y <- data[, yname]  
  
  if (normalize) {  
    y <- y - min(y)  
  }  
  
  if (is.null(xlabel)) {
```

```

    xlabel = xname
}
if (is.null(ylabel)) {
  ylabel = yname
}

aes_mapping <- aes(x = data[, xname], y = y)

data |>
  ggplot(mapping = aes_mapping) +
  ggtitle(title) +
  xlab(xlabel) +
  ylab(ylabel) +
  geom_point() +
  geom_line(color=line_col)
}

```

Here we also used `naive_model_MSEs` function defined in the `common-helper.functions.R` script (already mentioned above) to compute *Mean Squared Errors* using *5-Fold Cross Validation* method:

```

naive_model_MSEs <- function(val) {
  sapply(edx_CV, function(cv_item){
    mse(cv_item$validation_set$rating - val)
  })
}

```

One more function, defined in the `same script`, that we will need for further analysis of the current model, is the `naive_model_RMSE` one:

```

naive_model_RMSE <- function(val){
  sqrt(mean(naive_model_MSEs(val)))
}

```

2.3.1 Ensure that `mu.RMSE` value is the best for the current model

If we plug in any other number, we will get a higher RMSE. Let's prove that by the following small investigation:

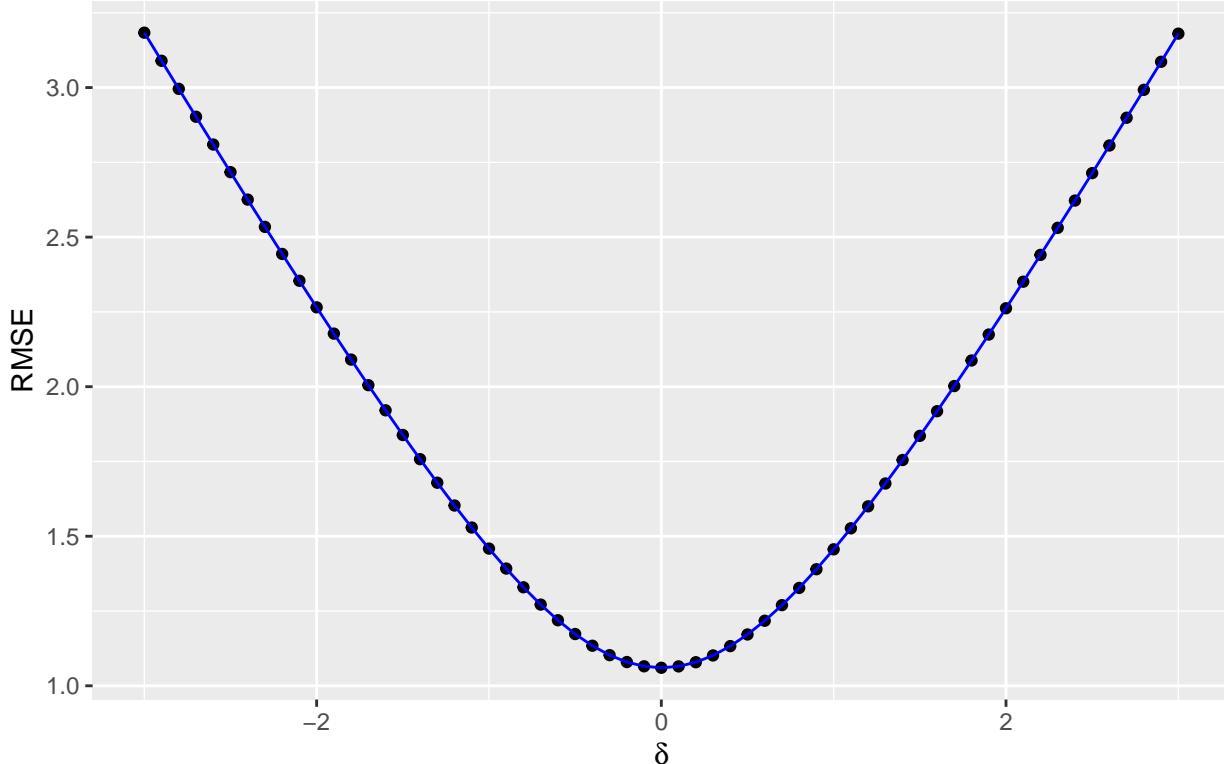
```
deviation <- seq(0, 6, 0.1) - 3

deviation.RMSE <- sapply(deviation, function(delta){
  naive_model_RMSE(mu + delta)
})
```

Let's make a quick investigation of the `deviation.RMSE` result we have just got:

```
data.frame(delta = deviation,
           delta.RMSE = deviation.RMSE) |>
  data.plot(title = TeX(r' [RMSE as a function of deviation ($\delta$) from the Overall Mean Rating ($\hat{\mu}$) ]'),
            xname = "delta",
            yname = "delta.RMSE",
            xlabel = TeX(r'[$\delta$]'),
            ylabel = "RMSE")
```

RMSE as a function of deviation (δ) from the Overall Mean Rating ($\hat{\mu}$)



```
which_min_deviation <- deviation[which.min(deviation.RMSE)]
min_rmse = min(deviation.RMSE)
```

```

print_log1("Minimum RMSE is achieved when the deviation from the mean is: %1",
          which_min_deviation)

## Minimum RMSE is achieved when the deviation from the mean is: 0

print_log1("Is the previously computed RMSE the best for the current model: %1",
          mu.RMSE == min_rmse)

## Is the previously computed RMSE the best for the current model: TRUE

RMSEs.ResultTibble.OMR <- RMSEs.ResultTibble |>
  RMSEs.AddRow("Overall Mean Rating Model", mu.RMSE)

RMSE_kable(RMSEs.ResultTibble.OMR)

```

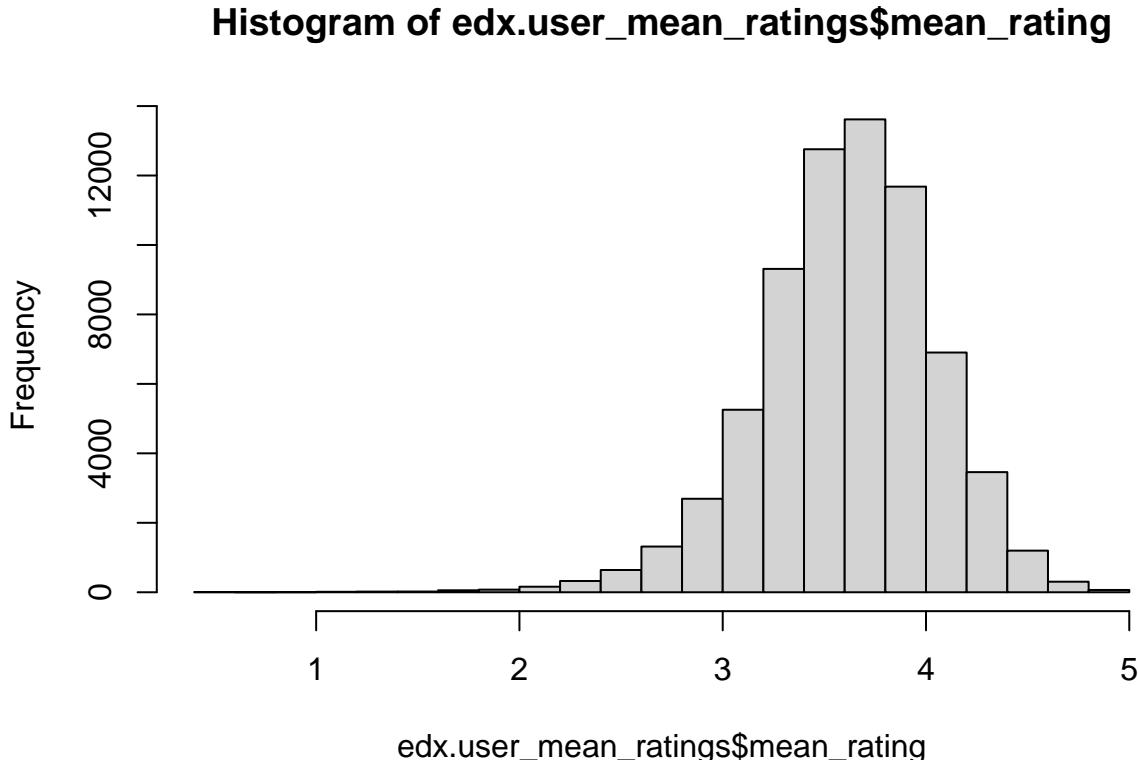
Method	RMSE	Comment
Project Objective	0.864900	
Overall Mean Rating Model	1.060346	

To win the grand prize of \$1,000,000, a participating team had to get an RMSE of at least 0.8563[2]. So we can definitely do better![8]

2.4 User Effect Model

To improve our model let's now take into consideration user effects as explained in [Section 23.4 User effects](#) of the *Course Textbook*. If we visualize the average rating for each user the way the [the author](#) shows, we can see that there is substantial variability in the average ratings across users:

```
hist(edx.user_mean_ratings$mean_rating, nclass = 30)
```



Following the author's further explanation, to account for this variability, we will use a linear model with a *treatment effect* α_i for each user. The sum $\mu + \alpha_i$ can be interpreted as the typical rating user i gives to movies. So we write the model as follows:

$$Y_{i,j} = \mu + \alpha_i + \varepsilon_{i,j}$$

Statistics textbooks refer to the α s as treatment effects. In the Netflix challenge papers, they refer to them as *bias*[9, 10].

As it is stated here[9], it can be shown that the least squares estimate $\hat{\alpha}_i$ is just the average of $y_{i,j} - \hat{\mu}$ for each user i . So we can compute them this way:

```
a <- rowMeans(y - mu, na.rm = TRUE)
```

These considerations allows us to compute a *User Mean Ratings* the following way:

```

put_log("Computing Average Ratings per User (User Mean Ratings)...")
user.mean_ratings <- rowMeans(edx.mx, na.rm = TRUE)
user_ratings.n <- rowSums(!is.na(edx.mx))

edx.user_mean_ratings <-
  data.frame(userId = names(user.mean_ratings),
             mean_rating = user.mean_ratings,
             n = user_ratings.n)

put_log("User Mean Ratings have been computed.")

str(edx.user_mean_ratings)

```

```

## 'data.frame':    69878 obs. of  3 variables:
## $ userId      : chr  "1" "2" "3" "4" ...
## $ mean_rating: num  5 3.29 3.94 4.06 3.92 ...
## $ n           : num  19 17 31 35 74 39 96 727 21 112 ...

```

And then we compute a *User Effect* this way:

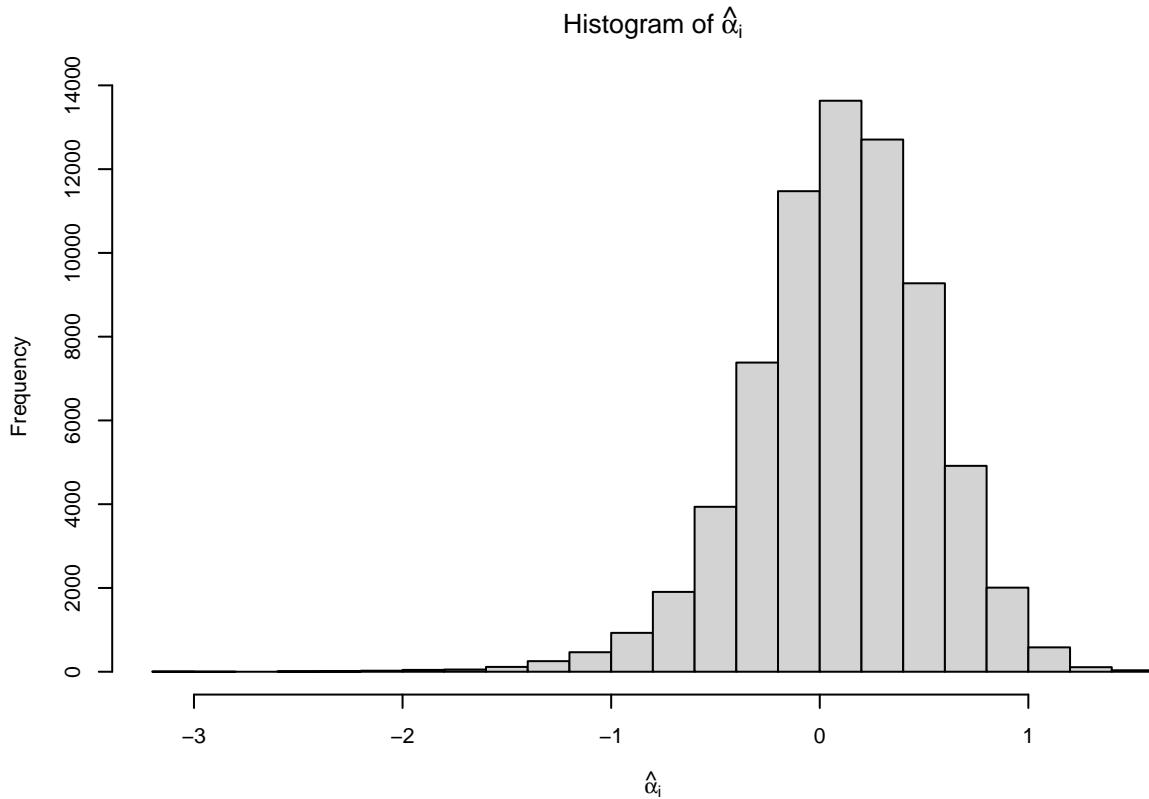
```

put_log("Computing User Effect per users ...")
edx.user_effect <- edx.user_mean_ratings |>
  mutate(userId = as.integer(userId),
         a = mean_rating - mu)

put_log("A User Effect Model has been builded")

par(cex = 0.7)
hist(edx.user_effect$a, 30, xlab = TeX(r'[\hat{\alpha}_i]'),
      main = TeX(r'[Histogram of \hat{\alpha}_i]'))

```



```
str(edx.user_effect)
```

```
## 'data.frame':   69878 obs. of  4 variables:
## $ userId      : int  1 2 3 4 5 6 7 8 9 10 ...
## $ mean_rating: num  5 3.29 3.94 4.06 3.92 ...
## $ n           : num  19 17 31 35 74 39 96 727 21 112 ...
## $ a           : num  1.488 -0.218 0.423 0.545 0.406 ...
```



The full source code of the *User Effect* computation is available in the [Model building: User Effect](#) section of the [capstone-movielens.main.R](#) script on *Github*.

Finally, we are ready to compute the RMSE (additionally using the `clamp` helper function we defined above to keep predictions in the proper range):

```
put_log("Computing the RMSE taking into account user effects...")
start <- put_start_date()
edx.user_effect.MSEs <- sapply(edx_CV, function(cv_fold_dat){
  cv_fold_dat$validation_set |>
    left_join(edx.user_effect, by = "userId") |>
    mutate(resid = rating - clamp(mu + a)) |>
    pull(resid) |> mse()
})
put_end_date(start)

edx.user_effect.RMSE <- sqrt(mean(edx.user_effect.MSEs))
```

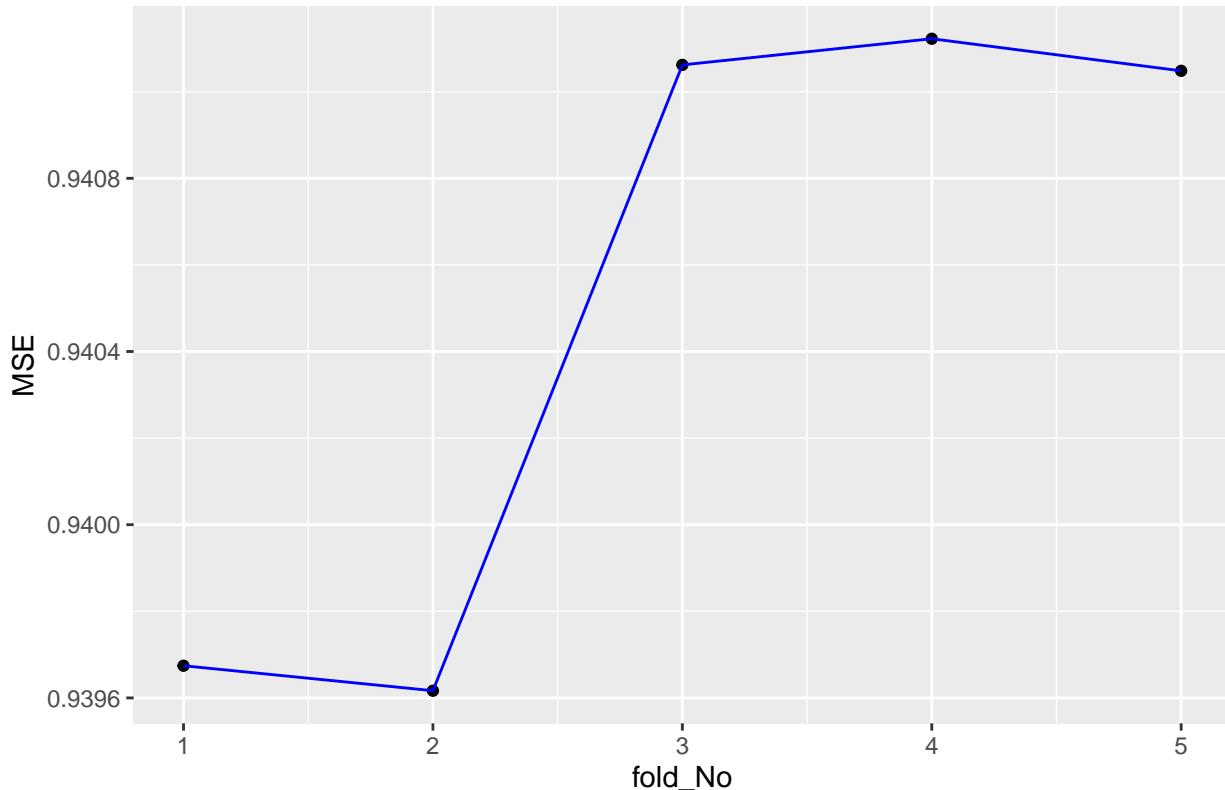
```

RMSEs.ResultTibble.UE <- RMSEs.ResultTibble.OMR |>
  RMSEs.AddRow("User Effect Model", edx.user_effect.RMSE)

data.frame(fold_No = 1:5, MSE = edx.user_effect.MSEs) |>
  data.plot(title = "MSE results of the 5-fold CV method applied to the User Effect Model",
            xlabel = "fold_No",
            ylabel = "MSE")

```

MSE results of the 5-fold CV method applied to the User Effect Model



```
RMSE_kable(RMSEs.ResultTibble.UE)
```

Method	RMSE	Comment
Project Objective	0.8649000	
Overall Mean Rating Model	1.0603462	
User Effect Model	0.9697962	



The full source code of the *User Effect Model RMSE* computation is available in the [Compute RMSE for User Effect Model](#) section of the [capstone-movielens.main.R](#) script on *GitHub*.

2.5 User+Movie Effect (UME) Model

In 23.5 Movie effects section of the *Course Textbook* the author draws our attention to the fact that some movies are generally rated higher than others. He also explains that a linear model with a *treatment effect* β_j for each movie can be used in this case, which can be interpreted as movie effect or the difference between the average ranking for movie j and the overall average μ :

$$Y_{i,j} = \mu + \alpha_i + \beta_j + \varepsilon_{i,j}$$

The author then shows how to use an approximation by first computing the least square estimate $\hat{\mu}$ and $\hat{\alpha}_i$, and then estimating $\hat{\beta}_j$ as the average of the residuals $y_{i,j} - \hat{\mu} - \hat{\alpha}_i$:

```
b <- colMeans(y - mu - a, na.rm = TRUE)
```

Inspired by this idea, a few support functions were developed by the author of this report, which we will use for our further analysis.

2.5.1 UME Model: Support Functions



The full source code of the functions described in this section is available in the [User+Movie Effect Model Functions](#) section of the [UM-effect.functions.R](#) script on *GitHub*.

2.5.1.1 `train_user_movie_effect` Function

We use this function to build and train our model using the `train_set` dataset:

```
train_user_movie_effect <- function(train_set, lambda = 0){  
  if (is.na(lambda)) {  
    stop("Function: train_user_movie_effect  
'lambda' is `NA`")  
  }  
  
  UM.effect <- train_set |>  
    left_join(edx.user_effect, by = "userId") |>  
    mutate(resid = rating - (mu + a)) |>  
    group_by(movieId) |>  
    summarise(b = mean_reg(resid, lambda), n = n())  
  
  stopifnot(!is.na(mean(UM.effect$b)))  
  UM.effect  
}
```



The function described above accepts the `lambda` parameter, which we will need later for the *Regularization* method. We also use the `mean_reg` function call, which we will also need later for the *Regularization* techniques (for details, see the `mean_reg` function description in the Section [Regularization: Common Helper Functions](#) section of the [Appendix](#) to this report). We will explain that later in the [Regularization Method] section. For now, we omit the `lambda` parameter, accepting its default value `lambda = 0`. In this case, the `mean_reg` function is equivalent to the standard R function `base::mean`.

2.5.1.2 train_user_movie_effect.cv Function

We use the `train_user_movie_effect.cv` function to build and train our model using the *5-Fold Cross Validation* method. Below, we provide the most important part of the code of that function:

```
train_user_movie_effect.cv <- function(lambda = 0){  
# ...  
  start <- put_start_date()  
  user_movie_effects_ls <- lapply(edx_CV, function(cv_fold_dat){  
    cv_fold_dat$train_set |> train_user_movie_effect(lambda)  
  })  
  put_end_date(start)  
  put_log("Function: train_user_movie_effect.cv:  
User+Movie Effect list have been computed")  
  
  user_movie_effects_united <- union_cv_results(user_movie_effects_ls)  
  
  user_movie_effect <- user_movie_effects_united |>  
    group_by(movieId) |>  
    summarise(b = mean(b), n = mean(n))  
  # ...  
  user_movie_effect  
}
```



Here we use the function call `union_cv_results`, which is defined in the script `common-helper.functions.R`, to aggregate the *5-Fold Cross Validation* method results (for details, see the `union_cv_results` function description in the [Appendix](#) of this report).

2.5.1.3 calc_user_movie_effect_MSE Function

The code of the function `calc_user_movie_effect_MSE` defined in the `UME.functions.R` script to calculate the Mean Squared Error (MSE) of the UME Model for the given Test Set is provided below:

```
calc_user_movie_effect_MSE <- function(test_set, um_effect){  
  mse.result <- test_set |>  
    left_join(edx.user_effect, by = "userId") |>  
    left_join(um_effect, by = "movieId") |>  
    mutate(resid = rating - clamp(mu + a + b)) |>  
    pull(resid) |> mse()  
  
  stopifnot(!is.na(mse.result))  
  mse.result  
}
```

2.5.1.4 calc_user_movie_effect_MSE.cv Function

The code of the function `calc_user_movie_effect_MSE.cv` defined in the `UME.functions.R` script to calculate the *5-Fold Cross Validation* MSE result of the UME Model is provided below:

```

calc_user_movie_effect_MSE.cv <- function(um_effect){
  put_log("Function: user_movie_effects_MSE.cv:
Computing the RMSE taking into account User+Movie Effects...")
  start <- put_start_date()
  user_movie_effects_MSEs <- sapply(edx_CV, function(cv_fold_dat){
    cv_fold_dat$validation_set |> calc_user_movie_effect_MSE(um_effect)
  })
  put_end_date(start)

  put_log1("Function: user_movie_effects_MSE.cv:
MSE values have been plotted for the %i-Fold Cross Validation samples.",
          CVFolds_N)

  mean(user_movie_effects_MSEs)
}

```

2.5.1.5 calc_user_movie_effect_RMSE Function

The code of the function `calc_user_movie_effect_RMSE` defined in the `UM-effect.functions.R` script to calculate the Root Mean Squared Error (RMSE) of the UME Model for the given Test Set is provided below:

```

calc_user_movie_effect_RMSE <- function(test_set, um_effect){
  mse <- test_set |> calc_user_movie_effect_MSE(um_effect)
  sqrt(mse)
}

```

2.5.1.6 calc_user_movie_effect_RMSE.cv Function

The code of the function `calc_user_movie_effect_RMSE.cv` defined in the `UM-effect.functions.R` script to calculate the 5-Fold Cross Validation RMSE result of the UME Model is provided below:

```

calc_user_movie_effect_RMSE.cv <- function(um_effect){
  user_movie_effects_MSE <- calc_user_movie_effect_MSE.cv(um_effect)
  um_effect_RMSE <- sqrt(user_movie_effects_MSE)
  put_log2("Function: user_movie_effects_RMSE.cv:
%1-Fold Cross Validation ultimate RMSE: %2", CVFolds_N, um_effect_RMSE)
  um_effect_RMSE
}

```

2.5.2 Model Building



The full source code of builing and training the current model is available in the [Model building: User+Movie Effect](#) section of the [capstone-movielens.main.R](#) script on [GitHub](#).

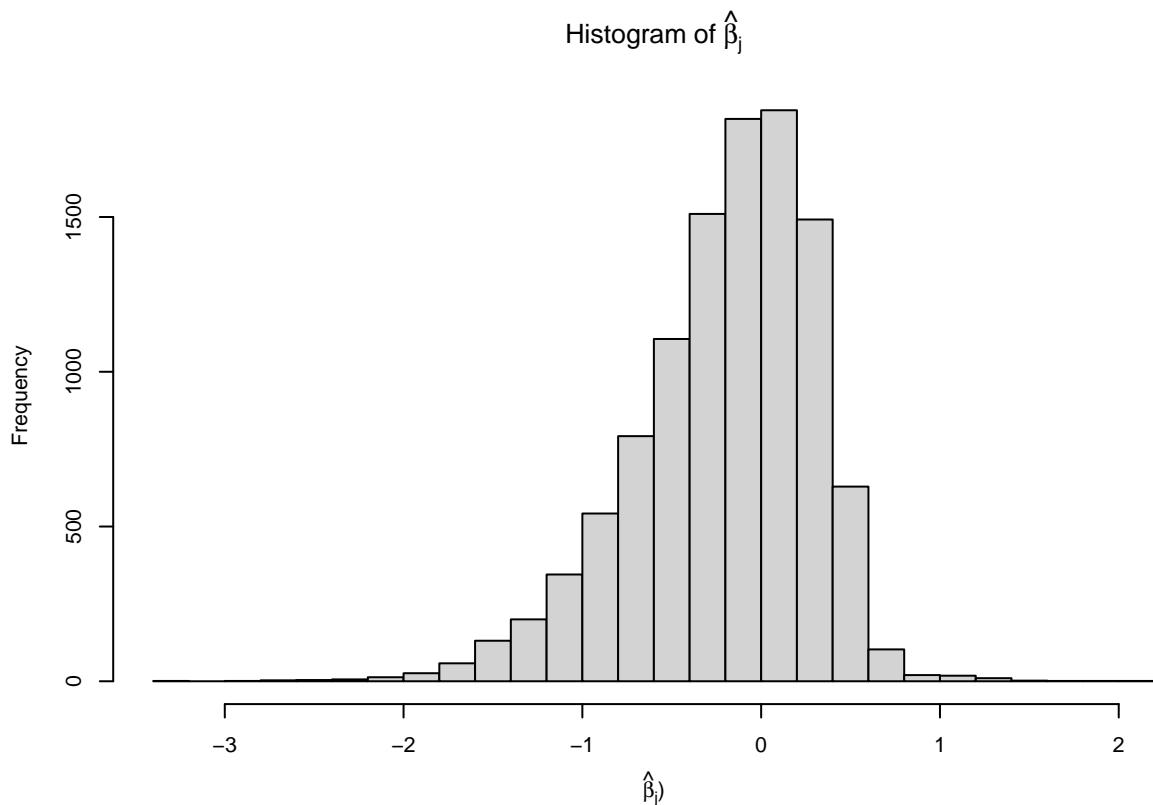
Below, we provide the most significant part of the code for training our model using the [5-Fold Cross Validation](#) method:

```
cv.UM_effect <- train_user_movie_effect.cv()

str(cv.UM_effect)

## # tibble [10,677 x 3] (S3: tbl_df/tbl/data.frame)
## $ movieId: int [1:10677] 1 2 3 4 5 6 7 8 9 10 ...
## $ b       : num [1:10677] 0.335 -0.306 -0.365 -0.598 -0.444 ...
## $ n       : num [1:10677] 18907 8593 5574 1253 5065 ...

par(cex = 0.7)
hist(cv.UM_effect$b, 30, xlab = TeX(r'[$\hat{\beta}_j$]'),
     main = TeX(r' [Histogram of $\hat{\beta}_j$]'))
```



We can now construct predictors and see how much the RMSE improves[11]:

```

cv.UM_effect.RMSE <- calc_user_movie_effect_RMSE.cv(cv.UM_effect)

RMSEs.ResultTibble.UME <- RMSEs.ResultTibble.UE |>
  RMSEs.AddRow("User+Movie Effect Model", cv.UM_effect.RMSE)

RMSE_kable(RMSEs.ResultTibble.UME)

```

Method	RMSE	Comment
Project Objective	0.8649000	
Overall Mean Rating Model	1.0603462	
User Effect Model	0.9697962	
User+Movie Effect Model	0.8732081	

2.5.3 User+Movie Effect Model Regularization

2.5.3.1 Utilizing Penalized least squares

Section 23.6 *Penalized least squares* of the *Course Textbook* explains why and how we should use *Penalized least squares* to improve our predictions. The author also explains that the general idea of penalized regression is to control the total variability of the movie effects: $\sum_{j=1}^n \beta_j^2$. Specifically, instead of minimizing the least squares equation, we minimize an equation that adds a penalty:

$$\sum_{i,j} (y_{u,i} - \mu - \alpha_i - \beta_j)^2 + \lambda \sum_j \beta_j^2$$

The first term is just the sum of squares and the second is a penalty that gets larger when many β_j s are large. Using calculus, we can actually show that the values of β_i that minimize this equation are:

$$\hat{\beta}_j(\lambda) = \frac{1}{\lambda + n_j} \sum_{i=1}^{n_j} (Y_{i,j} - \mu - \alpha_i)$$

where n_j is the number of ratings made for movie j .

This approach will have our desired effect: when our sample size n_j is very large, we obtain a stable estimate and the penalty λ is effectively ignored since $n_j + \lambda \approx n_j$. Yet when the n_j is small, then the estimate $\hat{\beta}_j(\lambda)$ is shrunken towards 0. The larger the λ , the more we shrink[12].

We will implement the *Regularization* method on our models (starting from the current model) in two steps:

1. **Preconfiguration:** Preliminary determination of the optimal range of λ values for the 5-Fold Cross Validation samples;
2. **Fine-tuning:** figuring out the value of λ that minimizes the model's RMSE.

2.5.3.2 UME Model Regularization: Support Functions



The `regularize.test_lambda.UM_effect.cv` function described below are defined in the Regularization section of the `UM-effect.functions.R` script.

2.5.3.2.1 `regularize.test_lambda.UM_effect.cv` Function

This function calculates *RMSE* of the *UME Model* using *5-Fold Cross Validation* method for the given λ parameter value:

```
regularize.test_lambda.UM_effect.cv <- function(lambda){  
  if (is.na(lambda)) {  
    stop("Function: regularize.test_lambda.UM_effect.cv  
`lambda` is `NA`")  
  }  
  um_effect <- train_user_movie_effect.cv(lambda)  
  calc_user_movie_effect_RMSE.cv(um_effect)  
}
```



Note that we reuse the function `train_user_movie_effect.cv` calling it from the `regularize.test_lambda.UM_effect.cv`, but now with the λ parameter different from the default (' $\lambda = 0$ ') value.

2.5.3.3 User+Movie Effect Model Tuning: Preconfiguration

Let's perform the preconfiguration to determine the appropriate range of λ for subsequent fine-tuning of our current model:

We are going to use the `tune.model_param` function described in the Appendix of this report, passing the `regularize.test_lambda.UM_effect.cv` function as the value of the `fn_tune.test.param_value` parameter:

```
lambdas <- seq(0, 1, 0.1)

cv.UME.preset.result <-
  tune.model_param(lambdas, regularize.test_lambda.UM_effect.cv)

put_log1("Preliminary regularization set-up of `lambda`'s range for the UME Model has been completed
for the %1-Fold Cross Validation samples.",
CVFolds_N)

str(cv.UME.preset.result)

## List of 2
## $ tuned.result:'data.frame':    8 obs. of  2 variables:
##   ..$ RMSE          : num [1:8] 0.873 0.873 0.873 0.873 ...
##   ..$ parameter.value: num [1:8] 0 0.1 0.2 0.3 0.4 0.5 0.6 0.7
## $ best_result : Named num [1:2] 0.4 0.873
##   .. - attr(*, "names")= chr [1:2] "param.best_value" "best_RMSE"
```

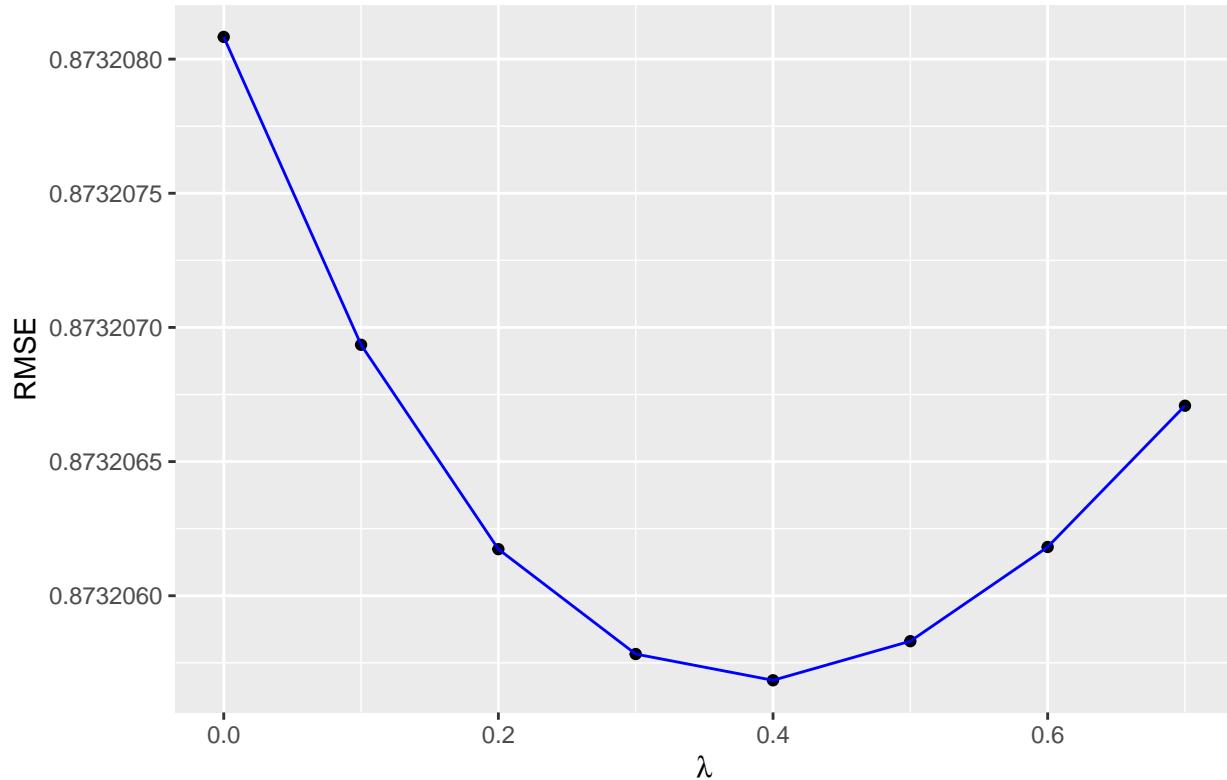


The complete version of the source code provided in this section can be found in the Process Preliminary setting-up of lambda range for UM Effect regularization section of the `capstone-movielens.main.R` script.

Now, let's visualize the results of the λ range preconfiguration:

```
cv.UME.preset.result$tuned.result |>
  data.plot(title = TeX(r'[Preliminary set-up of $\lambda$ for Regularazation of the User+Movie Effect Model]'),
            xname = "parameter.value",
            yname = "RMSE",
            xlabel = TeX(r'[$\lambda$]'),
            ylabel = "RMSE")
```

Preliminary set-up of λ for Regulararazation of the User+Movie Effect M



We use the custom data visualization function `data.plot` described in the [Appendix](#) of this report, which is defined in the [Data Visualization](#) section of the `data.helper.functions.R` script.

2.5.3.4 User+Movie Effect Model: Fine-Tuning

2.6 User+Movie+Genre Effect (UMGE) Model

As mentioned in [Section 23.7: Exercises](#) of the *Chapter “23 Regularization” of the Course Textbook* the `Movielens` dataset also has a genres column. This column includes every genre that applies to the movie (some movies fall under several genres)[[13](#)].

2.6.1 Genre Data Analysis

```
#### Movie Genres Data
```

The following code computes movie rating summaries by popular genres like Drama, Comedy, Thriller, and Romance:

```
#library(stringr)
genres = c("Drama", "Comedy", "Thriller", "Romance")
sapply(genres, function(g) {
  sum(str_detect(edx$genres, g))
})
```

Further, we can find out the movies that have the greatest number of ratings using the following code:

```
ordered_movie_ratings <- edx |> group_by(movieId, title) |>
  summarise(number_of_ratings = n()) |>
  arrange(desc(number_of_ratings))
print(head(ordered_movie_ratings))
```

and figure out the most given ratings in order from most to least:

```
ratings <- edx |> group_by(rating) |>
  summarise(count = n()) |>
  arrange(desc(count))
print(ratings)
```

The following code allows us to summarize that in general, half-star ratings are less common than whole-star ratings (e.g., there are fewer ratings of 3.5 than there are ratings of 3 or 4, etc.):

```
print(edx |> group_by(rating) |> summarise(count = n()))
```

We can visually see that from the following plot:

```
edx |>
  group_by(rating) |>
  summarise(count = n()) |>
  ggplot(aes(x = rating, y = count)) +
  geom_line()
```

2.6.1.1 Movie Genres Effect

The plot below shows strong evidence of a genre effect (for illustrative purposes, the plot shows only categories with more than 20, 000 ratings).

```
# Preparing data for plotting:
genre_ratins_grp <- train_set |>
  mutate(genre_categories = as.factor(genres)) |>
  group_by(genre_categories) |>
  summarize(n = n(), rating_avg = mean(rating), se = sd(rating)/sqrt(n())) |>
  filter(n > 20000) |>
  mutate(genres = reorder(genre_categories, rating_avg)) |>
  select(genres, rating_avg, se, n)

dim(genre_ratins_grp)
genre_ratins_grp_sorted <- genre_ratins_grp |> sort_by.data.frame(~ rating_avg)
print(genre_ratins_grp_sorted)

# Creating plot:
genre_ratins_grp |>
  ggplot(aes(x = genres, y = rating_avg, ymin = rating_avg - 2*se, ymax = rating_avg + 2*se)) +
  geom_point() +
  geom_errorbar() +
  ggtitle("Average rating per Genre") +
  ylab("Average rating") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```

Below are worst and best ratings categories:

```
sprintf("The worst ratings are for the genre category: %s",
       genre_ratins_grp$genres[which.min(genre_ratins_grp$genres)])

sprintf("The best ratings are for the genre category: %s",
       genre_ratins_grp$genres[which.max(genre_ratins_grp$genres)])
```

Another way of visualizing a genre effect is shown in the section [Average rating for each genre](#) of the article “Movie Recommendation System using R - BEST” written by [Amir Moterfaker](#)[6]:

```
# For better visibility, we reduce the data for plotting
# while keeping the worst and best rating rows:
plot_ind <- odd(1:nrow(genre_ratins_grp))
plot_dat <- genre_ratins_grp_sorted[plot_ind,]

plot_dat |>
  ggplot(aes(x = rating_avg, y = genres)) +
  ggtitle("Genre Average Rating") +
  geom_bar(stat = "identity", width = 0.6, fill = "#8888ff") +
  xlab("Average ratings") +
  ylab("Genres") +
  scale_x_continuous(labels = comma, limits = c(0.0, 5.0)) +
  theme_economist() +
  theme(plot.title = element_text(vjust = 3.5),
        axis.title.x = element_text(vjust = -5, face = "bold"),
```

```

axis.title.y = element_text(vjust = 10, face = "bold"),
axis.text.x = element_text(vjust = 1, hjust = 1, angle = 0),
axis.text.y = element_text(vjust = 0.25, hjust = 1, size = 8),
plot.margin = margin(0.7, 0.5, 1, 1.2, "cm"))

```

If we define $g_{i,j}$ as the genre for user's i rating of movie j , we can use the following models to account for the **genre** effect:

To account for *genre effects* we will use the model suggested in the [Section 23.7: Exercises](#) of the *Chapter “23 Regularization” of the Course Textbook*[[13](#)]:

$$Y_{i,j} = \mu + \alpha_i + \beta_j + g_{i,j} + \varepsilon_{i,j}$$

where $g_{i,j}$ is an *aggregation function* which is explained in detail in *Section 22.3: “Review of Aggregation Functions” of “Recommender Systems Handbook” (Chapter 22: “Aggregation of Preferences in Recommender Systems”, p. 712)* book[[14](#)].

In the formula above $g_{i,j}$ denotes a *genre effect* for user's i rating of movie j , so that:

$$g_{i,j} = \sum_{k=1}^K x_{i,j}^k \gamma_k$$

with $x_{i,j}^k = 1$ if $g_{i,j}$ includes genre k , and $x_{i,j}^k = 0$ otherwise.

$$Y_{i,j} = \mu + \alpha_i + \beta_j + g_{i,j} + f(d_{i,j})$$

$$\sum_{i=1}^{n_i} (Y_{i,j} - \mu - \alpha_i)$$

2.7 User+Movie+Genre+Year Effect (UMGYE) Model

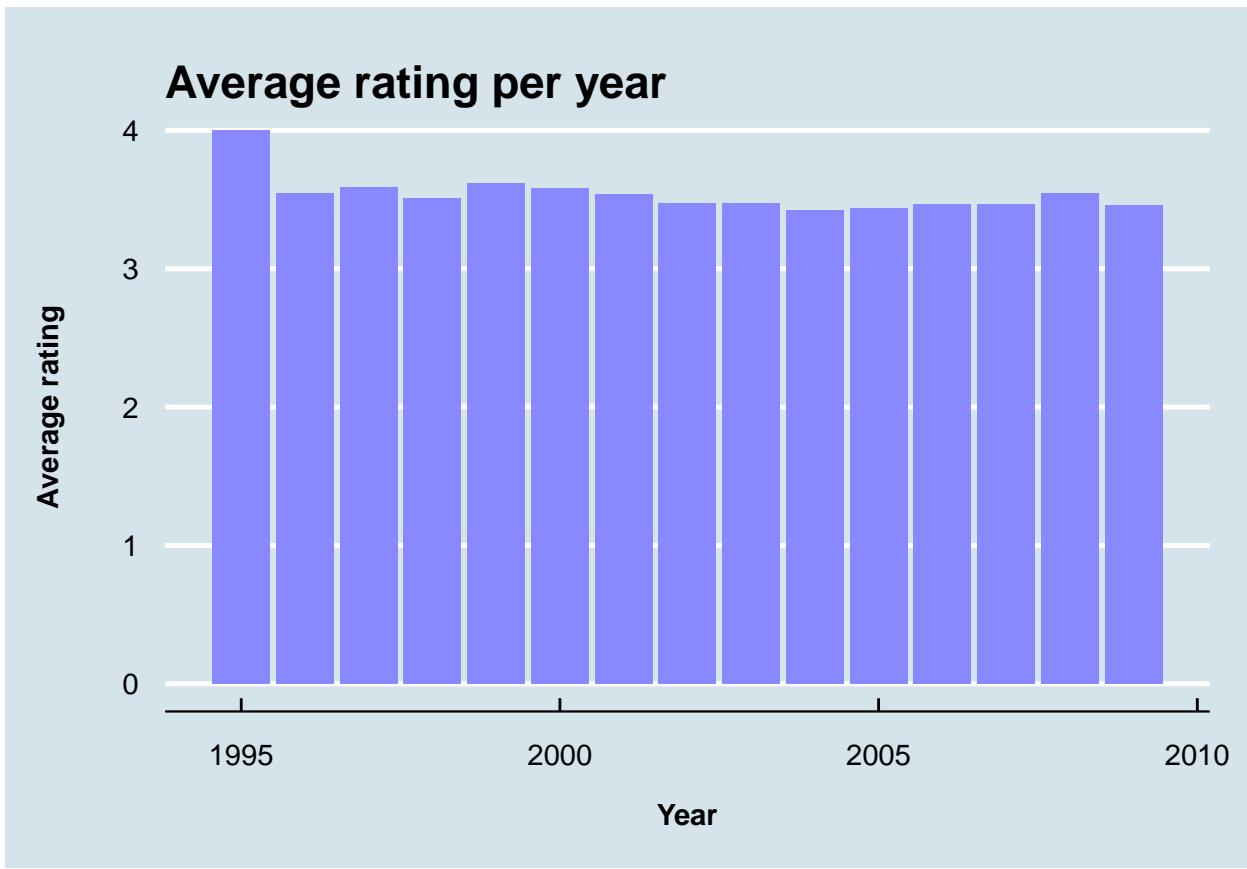
2.7.0.1 Yearly rating count[6]

```
print(edx |>
  mutate(year = year(as_datetime(timestamp, origin = "1970-01-01")))) |>
  group_by(year) |>
  summarize(count = n())
)

## # A tibble: 15 x 2
##       year   count
##     <dbl>   <int>
## 1    1995      2
## 2    1996  942772
## 3    1997  414101
## 4    1998 181634
## 5    1999  709893
## 6    2000 1144349
## 7    2001  683355
## 8    2002  524959
## 9    2003  619938
## 10   2004  691429
## 11   2005 1059277
## 12   2006  689315
## 13   2007  629168
## 14   2008  696740
## 15   2009  13123
```

2.7.0.2 Average rating per year plot[6]

```
edx |>
  mutate(year = year(as_datetime(timestamp, origin = "1970-01-01")))) |>
  group_by(year) |>
  summarize(rating_avg = mean(rating)) |>
  ggplot(aes(x = year, y = rating_avg)) +
  geom_bar(stat = "identity", fill = "#8888ff") +
  ggtitle("Average rating per year") +
  xlab("Year") +
  ylab("Average rating") +
  scale_y_continuous(labels = comma) +
  theme_economist() +
  theme(axis.title.x = element_text(vjust = -5, face = "bold"),
        axis.title.y = element_text(vjust = 10, face = "bold"),
        plot.margin = margin(0.7, 0.5, 1, 1.2, "cm"))
```



We use the following models to account for the `date` effect:

$$Y_{i,j} = \mu + \alpha_i + \beta_j + f(d_{i,j}) + \varepsilon_{i,j}$$

2.8 User+Movie+Genre+Year+Day Effect (UMGYDE) Model

2.8.1 UMGYDE Model: Matrix Factorization

3 Conclusion

Hello Conclusion!

This is a great conclusion, isn't it?!!

4 Appendix

4.1 Data Helper Functions

4.1.1 `union_cv_results` Function

Aggregates the input data frames into a single data frame.

4.1.1.1 Usage

```
union_cv_results(data_list = data)
```

4.1.1.2 Parameters

::: { #func.mean_reg.params.data_list .sidebar } - **data_list**: List of data frames representing the results of the *N-Fold Cross Validation* method execution.;

:::

4.1.1.3 Details

The function is used to aggregate the *N-Fold Cross Validation* method result data into a single data frame for further processing.

4.1.1.4 Value

A data frame that is a union of the input data frames.

4.1.1.5 Source Code

The source code of the `union_cv_results` function is shown below:

```
union_cv_results <- function(data_list) {
  out_dat <- data_list[[1]]

  for (i in 2:CVFolds_N){
    out_dat <- union(out_dat,
                      data_list[[i]])
  }

  out_dat
}
```



The source code of the `union_cv_results` is also available in the [Model training](#) section of the `data.helper.functions.R` script.

4.1.2 `data.plot` Function

4.1.2.1 Usage

```
data.plot(data,
          title,
          xname,
          yname,
          xlabel = NULL,
          ylabel = NULL,
          line_col = "blue",
          # scale = 1,
          normalize = FALSE)
```

4.1.2.2 Parameters

::: { #func.mean_reg.params.param1 .sidebar } - **param1**: description;

- **param2**: description.

:::

4.1.2.3 Details



Note

4.1.2.4 Value

4.1.2.5 Source Code

The soure code of the `data.plot` function is shown below:

```
data.plot <- function(data,
                      title,
                      xname,
                      yname,
                      xlabel = NULL,
```

```

    ylabel = NULL,
    line_col = "blue",
    # scale = 1,
    normalize = FALSE) {

y <- data[, yname]

if (normalize) {
  y <- y - min(y)
}

if (is.null(xlabel)) {
  xlabel = xname
}
if (is.null(ylabel)) {
  ylabel = yname
}

aes_mapping <- aes(x = data[, xname], y = y)

data |>
  ggplot(mapping = aes_mapping) +
  ggtitle(title) +
  xlab(xlabel) +
  ylab(ylabel) +
  geom_point() +
  geom_line(color=line_col)
}

```



The source code of the `data.plot` is also available in the [Data Visualization](#) section of the `data.helper.functions.R` script.

4.2 Regularization: Common Helper Functions



The full source code of the functions described below are available in the [Model Tuning](#) section of the `common-helper.functions.R` script on [GitHub](#).

4.2.1 `mean_reg` Function

A general function for computing the arithmetic mean given a *regularization parameter* λ . We will call it the *penalized mean*.

4.2.1.1 Usage

```
mean_reg(vals, lambda = 0, na.rm = TRUE)
```

4.2.1.2 Parameters

- **vals:** values (usually a numeric vector) to calculate the *penalized mean*;
- **lambda:** a *regularization parameter* λ used in the *Regularization techniques*.
- **na.rm:** a logical evaluating to TRUE or FALSE indicating whether NA values should be stripped before the computation proceeds.

4.2.1.3 Details

We call this function internally from the functions that compute the *penalized estimates* as described in section [23.6 Penalized Least Squares](#) of the *Course Textbook*[12].



If the `lambda` parameter is 0 (the default), the function is equivalent to the standard R function `base::mean`, calling with the parameter `trim = 0` (the default for the `base::mean` function).

4.2.1.4 Value

The function calculates the *penalized mean* of the `vals` parameter as follows:

$$\mu(\lambda) = \frac{1}{\lambda + N} \sum_{i=1}^N x_i$$

When the `lambda` parameter is 0 (meaning the $\lambda = 0$), the function calculates the simple arithmetic mean as follows:

$$\mu = \frac{1}{N} \sum_{i=1}^N x_i$$

4.2.1.5 Source Code

The source code of the `mean_reg` function is shown below:

```

mean_reg <- function(vals, lambda = 0, na.rm = TRUE){
  if (is.na(lambda)) {
    stop("Function: mean_reg
`lambda` is `NA`")
  }

  names(lambda) <- NULL
  sums <- sum(vals, na.rm = na.rm)
  N <- ifelse(na.rm, sum(!is.na(vals)), length(vals))
  sums/(N + lambda)
}

```



The source code of the `mean_reg` is also available in the [Regularization](#) section of the [common-helper.functions.R](#) script.

4.2.2 `tune.model_param` Function

The function searches for the parameter value corresponding to the minimum value of the RMSE from the list of values specified by the `param_values` parameter.

4.2.2.1 Signature

```

tune.model_param <- function(param_values,
                               fn_tune.test.param_value,
                               break.if_min = TRUE,
                               steps.beyond_min = 2){

  # ...
  list(tuned.result = data.frame(RMSE = RMSEs_tmp,
                                 parameter.value = param_vals_tmp),
       best_result = param_values.best_result)
}

```

4.2.2.2 Parameters

- **`param_values`:** A list of values to search for the value corresponding to the minimum value of the RMSE ;
- **`fn_tune.test.param_value`:** A helper function that calculates the value of the RMSE for a given parameter value.;
- **`break.if_min = TRUE`:** A Boolean parameter that determines whether the function should terminate after completing the number of steps specified by the parameter `steps.beyond_min`, after the minimum value of the RMSE has been found;

- **steps.beyond_min = 2:** (takes effect only if `break.if_min` parameter is TRUE) Specifies the number of steps after finding the minimum value of the RMSE, upon completion of which the function should terminate.

4.2.2.3 Details

During execution, the function uses a helper function specified by the `fn_tune.test.param_value` parameter, which calculates the RMSE value for the given parameter from the list determined by the `param_values` parameter.



Note that the algorithm assumes that the dependence of the RMSE on the input parameter is a monotonically decreasing function until a minimum is reached and monotonically increasing thereafter. That is, it is assumed that the function has a single minimum on the given interval.

4.2.2.4 Value

The function returns a data structure containing the found value of the input parameter `param_values` for which the RMSE value is minimal, as well as the minimum RMSE value itself, along with a sequence of all calculated RMSE values:

```
list(tuned.result = data.frame(RMSE = RMSEs_tmp,
                                parameter.value = param_vals_tmp),
     best_result = param_values.best_result)
```

4.2.2.5 Source Code

Below is the most significant part of the source code of the `tune.model_param` function:

```
tune.model_param <- function(param_values,
                               fn_tune.test.param_value,
                               break.if_min = TRUE,
                               steps.beyond_min = 2){
  n <- length(param_values)
  param_vals_tmp <- numeric()
  RMSEs_tmp <- numeric()
  RMSE_min <- Inf
  i_max.beyond_RMSE_min <- Inf
  prm_val.best <- NA

  # ...

  for (i in 1:n) {
    put_log1("Function: `tune.model_param` :
Iteration %1", i)
    prm_val <- param_values[i]
    param_vals_tmp[i] <- prm_val

    RMSE_tmp <- fn_tune.test.param_value(prm_val)
    RMSEs_tmp[i] <- RMSE_tmp

    plot(param_vals_tmp[RMSEs_tmp], RMSEs_tmp[RMSEs_tmp])

    if(RMSE_tmp > RMSE_min){
```

```

    warning("Function: `tune.model_param`:
`RSME` reached its minimum: ", RMSE_min, "
for parameter value: ", prm_val)
    put_log2("Function: `tune.model_param`:
Current `RMSE` value is %1 related to parameter value: %2",
          RMSE_tmp,
          prm_val)

    if (i > i_max.beyond_RMSE_min) {
        warning("Function: `tune.model_param`:
Operation is broken (after `RSME` reached its minimum) on the following step: ", i)
        break
    }
    next
}

RMSE_min <- RMSE_tmp
prm_val.best <- prm_val

if (break.if_min) {
    i_max.beyond_RMSE_min <- i + steps.beyond_min
}
}

param_values.best_result <- c(param.best_value = prm_val.best,
                               best_RMSE = RMSE_min)

list(tuned.result = data.frame(RMSE = RMSEs_tmp,
                               parameter.value = param_vals_tmp),
     best_result = param_values.best_result)
}

```

4.2.3 `model.tune.param_range` Function

The function fine-tunes the model by searching for the best possible value of the input parameter over a given interval for which the corresponding RMSE value is minimal.

4.2.3.1 Signature

```
model.tune.param_range <- function(loop_starter,
                                      tune_dir_path,
                                      cache_file_base_name,
                                      fn_tune.test.param_value,
                                      max.identical.min_RMSE.count = 4,
                                      endpoint.min_diff = 0,
                                      break.if_min = TRUE,
                                      steps.beyond_min = 2){
  # ...
  list(best_result = param_values.best_result,
       param_values.endpoints = c(prm_val.leftmost, prm_val.rightmost, seq_increment),
       tuned.result = data.frame(parameter.value = parameter.value,
                                  RMSE = result.RMSE))
}
```

4.2.3.2 Parameters

4.2.3.2.1 *loop_starter*

A numeric vector of the form `c(start, end, dvs)`, where `start` and `end` are the endpoints of the interval on which the parameter value that minimizes RMSE is sought. `dvs` is a divisor for splitting the interval to transform it into a sequence of values among which the value that minimizes RMSE is sought. For this purpose, the sequence step is calculated as follows:

$$step = \frac{end - start}{dvs}$$

The sequence obtained as a result of the transformation is equivalent to the one generated by the function `seq` as follows:

```
seq(start, end, step)
```

In fact, the `seq` function is called internally to generate the sequence during the execution of the `model.tune.param_range` function.

4.2.3.2.2 *tune_dir_path*

To improve performance, the algorithm caches intermediate results in the file system. This parameter specifies the path to the directory where the files are cached.

4.2.3.2.3 *cache_file_base_name*

The algorithm generates unique names for cache files based on this and the `loop_starter` parameter, as well as some other intermediate values calculated during the execution.

4.2.3.2.4 *fn_tune.test.param_value*

This is a helper function name that is passed to the same-named parameter of the `tune.model_param` function that is called internally during the execution (see the description of the `tune.model_param` function [above](#)).

4.2.3.2.5 *max.identical.min_RMSE.count = 4*

If more than one identical minimum RMSE value is calculated during execution, the number of identical minimums is limited by the value of this parameter. When it is reached, the algorithm considers the task execution to be complete.

4.2.3.2.6 *endpoint.min_diff = 0*

Defines the sensitivity threshold for determining the neighborhood boundaries of the minimum RMSE value (for details, see the **Details** section [below](#)).

4.2.3.2.7 *break.if_min = TRUE*

This is a parameter that is required for the `tune.model_param` function that is called internally during execution (see the description of the `tune.model_param` function [above](#)).

4.2.3.2.8 *steps.beyond_min = 2*

This is a parameter that is required for the `tune.model_param` function that is called internally during execution (see the description of the `tune.model_param` function [above](#)).

4.2.3.3 Details

First, the function generates a sequence of the `input parameter values` based on the `loop_starter` parameter values, as described above (see the `loop_starter` parameter description for the details), which is used as one of the input parameters for the helper function `tune.model_param`, which is repeatedly called during the execution, performing fine-tuning of the model.

The `tune.model_param` function returns a range of input parameter values associated with the set of corresponding *Root Mean Squared Errors* that is also guaranteed to include their minimum value, as described in the **Value** subsection of the `tune.model_param` function description.

Next, the function figures out the boundary indices of a range of values from the neighborhood of the minimum RMSE by calling internally another helper function `get_fine_tune.param.endpoints.idx` (see the description of the function `get_fine_tune.param.endpoints.idx` below) and, using one more helper function `get_best_param.result` (see the description of the function `get_best_param.result` below), a pair of values, corresponding to the best result: minimal RMSE and corresponding input parameter value (which is considered the best).

The found values of the boundary indices are then used as the endpoints of a new interval to regenerate the sequence based on it in the next iteration, just as it was done based on the values of the `loop_starter` parameter at the very beginning of the execution. The value of `step divisor`, used to calculate the step of the generated sequence, remains unchanged during the entire execution (see the description of the `loop_starter` parameter above for details).

Thus, with each subsequent iteration, the minimum RMSE value is calculated more accurately, over an ever-decreasing interval, with the boundary values tending to the minimum RMSE value, and the sequence generated with an ever-decreasing step.

The calculation is completed when subsequent calculated values of the minimum RMSE stop improving and reach the most accurate possible value.

4.2.3.4 Value

A data structure containing the minimum RMSE value reached during the fine-tuning process, the corresponding input parameter value, and information about the final sequence, on which the best output values were found:

```
list(best_result = param_values.best_result,
     param_values.endpoints = c(prm_val.leftmost, prm_val.rightmost, seq_increment),
     tuned.result = data.frame(parameter.value = parameter.value,
                               RMSE = result.RMSE))
```

4.2.3.5 Source Code

Below is the simplified version of the source code of the `model.tune.param_range` function:

```
model.tune.param_range <- function(loop_starter,
                                      tune_dir_path,
                                      cache_file_base_name,
                                      fn_tune.test.param_value,
                                      max.identical.min_RMSE.count = 4,
                                      is.cv = TRUE,
                                      endpoint.min_diff = 0, #1e-07,
                                      break.if_min = TRUE,
                                      steps.beyond_min = 2){

  seq_start <- loop_starter[1]
  seq_end <- loop_starter[2]
  interval_divisor <- loop_starter[3]

  if (interval_divisor < 4) {
    interval_divisor <- 4
  }

  prm_val.leftmost <- seq_start
  prm_val.rightmost <- seq_end

  RMSE.leftmost <- NA
  RMSE.rightmost <- NA

  best_RMSE <- NA
  param.best_value <- 0

  param_values.best_result <- c(param.best_value = param.best_value,
                                 best_RMSE = best_RMSE)
  # Start repeat loop
  repeat{
    seq_increment <- (seq_end - seq_start)/interval_divisor

    if (seq_increment < 0.00000000000001) {
      warning("Function `model.tune.param_range`:
parameter value increment is too small.")
      break
    }
  }
}
```

```

}

test_param_vals <- seq(seq_start, seq_end, seq_increment)

tuned_result <- tune.model_param(test_param_vals,
                                    fn_tune.test.param_value,
                                    break.if_min,
                                    steps.beyond_min)

tuned.result <- tuned_result$tuned.result
plot(tuned.result$parameter.value, tuned.result$RMSE)

bound.idx <- get_fine_tune.param.endpoints.idx(tuned.result)
start.idx <- bound.idx["start"]
end.idx <- bound.idx["end"]
best_RMSE.idx <- bound.idx["best"]

prm_val.leftmost.tmp <- tuned.result$parameter.value[start.idx]
RMSE.leftmost.tmp <- tuned.result$RMSE[start.idx]

prm_val.rightmost.tmp <- tuned.result$parameter.value[end.idx]
RMSE.rightmost.tmp <- tuned.result$RMSE[end.idx]

min_RMSE <- tuned.result$RMSE[best_RMSE.idx]
min_RMSE.prm_val <- tuned.result$parameter.value[best_RMSE.idx]

seq_start <- prm_val.leftmost.tmp
seq_end <- prm_val.rightmost.tmp

if (is.na(best_RMSE)) {
  prm_val.leftmost <- prm_val.leftmost.tmp
  RMSE.leftmost <- RMSE.leftmost.tmp

  prm_val.rightmost <- prm_val.rightmost.tmp
  RMSE.rightmost <- RMSE.rightmost.tmp

  param.best_value <- min_RMSE.prm_val
  best_RMSE <- min_RMSE
}

if (RMSE.leftmost.tmp - min_RMSE >= endpoint.min_diff) {
  prm_val.leftmost <- prm_val.leftmost.tmp
  RMSE.leftmost <- RMSE.leftmost.tmp
}

if (RMSE.rightmost.tmp - min_RMSE >= endpoint.min_diff) {
  prm_val.rightmost <- prm_val.rightmost.tmp
  RMSE.rightmost <- RMSE.rightmost.tmp
}

if (end.idx - start.idx <= 0) {
  warning(`tuned.result$parameter.value` sequential start index are the same or greater than end or
break

```

```

}

if (best_RMSE == min_RMSE) {
  warning("Currently computed minimal RMSE equals the previously reached best one: ",
         best_RMSE,
  Currently computed minial value is: ", min_RMSE)

  if (sum(tuned.result$RMSE[tuned.result$RMSE == min_RMSE]) >= max.identical.min_RMSE.count) {
    warning("Minimal `RMSE` identical values count reached it maximum allowed value: ",
           max.identical.min_RMSE.count)

    param_values.best_result <-
      get_best_param.result(tuned.result$parameter.value,
                            tuned.result$RMSE)
    break
  }

} else if (best_RMSE < min_RMSE) {
  stop("Current minimal RMSE is greater than previously computed best value: ",
       best_RMSE,
  Currently computed minial value is: ", min_RMSE)
}

best_RMSE <- min_RMSE
param.best_value <- min_RMSE.prm_val

param_values.best_result <-
  get_best_param.result(tuned.result$parameter.value,
                        tuned.result$RMSE)
}

# End repeat loop

n <- length(tuned.result$parameter.value)
parameter.value <- tuned.result$parameter.value
result.RMSE <- tuned.result$RMSE

if (result.RMSE[1] == best_RMSE) {
  parameter.value[1] <- prm_val.leftmost
  result.RMSE[1] <- RMSE.leftmost
}
if (result.RMSE[n] == best_RMSE) {
  parameter.value[n+1] <- prm_val.rightmost
  result.RMSE[n+1] <- RMSE.rightmost
}

list(best_result = param_values.best_result,
     param_values.endpoints = c(prm_val.leftmost, prm_val.rightmost, seq_increment),
     tuned.result = data.frame(parameter.value = parameter.value,
                               RMSE = result.RMSE))
}

```



The full version of the source code of the `model.tune.param_range` is available in the [Model Tuning](#) section of the `common-helper.functions.R` script.

4.2.4 `get_fine_tune.param.endpoints.idx` Function

4.2.4.1 Signature

```
get_fine_tune.param.endpoints.idx <- function(preset.result) {  
  # ...  
  
  c(start = i,  
    end = j,  
    best = best_RMSE.idx)  
}
```

4.2.4.2 Parameters

4.2.4.2.1 *preset.result*

A data structure compatible with the `tuned.result` item of the data structure returned by the `tune.model_param` function (see the `Value` subsection of the `tune.model_param` function description section for more details). Here is a sample of using this parameter in the `model.tune.param_range` function code:

```
tuned_result <- tune.model_param(test_param_vals,  
                                    fn_tune.test.param_value,  
                                    break.if_min,  
                                    steps.beyond_min)  
  
tuned.result <- tuned_result$tuned.result  
# tuned.result = data.frame(RMSE, parameter.value)  
# ...  
bound.idx <- get_fine_tune.param.endpoints.idx(tuned.result)  
# boundnd.idx = c(start.index, end.index, min_RMSE.index)
```

4.2.4.3 Details

The algorithm finds the indices of the nearest neighboring elements of the element corresponding to the minimum RMSE value in the given sequence.

4.2.4.4 Value

A vector containing the index of the element corresponding to the minimum RMSE value in a given sequence, along with the indices of its nearest neighboring elements:

```
c(start, # interval start index  
  end,   # interval end index  
  best)  # index of the best `RMSE`
```

4.2.4.5 Source Code

The source code of the `get_fine_tune.param.endpoints.idx` function is shown below:

```
get_fine_tune.param.endpoints.idx <- function(preset.result) {  
  best_RMSE <- min(preset.result$RMSE)  
  best_RMSE.idx <- which.min(preset.result$RMSE)  
  # best_lambda <- preset.result$parameter.value[best_RMSE.idx]  
  
  preset.result.N <- length(preset.result$RMSE)  
  i <- best_RMSE.idx  
  j <- i  
  
  while (i > 1) {  
    i <- i - 1  
  
    if (preset.result$RMSE[i] > best_RMSE) {  
      break  
    }  
  }  
  
  while (j < preset.result.N) {  
    j <- j + 1  
  
    if (preset.result$RMSE[j] > best_RMSE) {  
      break  
    }  
  }  
  
  c(start = i,  
    end = j,  
    best = best_RMSE.idx)  
}
```

4.2.5 `get_best_param.result` Function

4.2.5.1 Signature

```
get_best_param.result <- function(param_values, rmses){  
  # ...  
  
  c(param.best_value = param_values[best_pvalue_idx],  
    best_RMSE = rmses[best_pvalue_idx])
```

4.2.5.2 Parameters

A pair of matched vectors, the first of which contains the values of the input parameters for tuning, and the second, the corresponding values of the RMSE.

- **param_values:** A vector containing the values of the input parameters for tuning;;
- **rmses:** A vector containing values of the corresponding Root Mean Squared Errors..

4.2.5.3 Details

Extracts the best values from the input data and returns them as a vector containing a pair of the best *input parameter* value and the corresponding *minimum RMSE*.

4.2.5.4 Value

A vector containing a pair of the best *input parameter* value and the corresponding *minimum RMSE*:

```
c(param.best_value, best_RMSE)
```

4.2.5.5 Source Code

The source code of the `get_best_param.result` function is shown below:

```
get_best_param.result <- function(param_values, rmses){  
  best_pvalue_idx <- which.min(rmses)  
  c(param.best_value = param_values[best_pvalue_idx],  
    best_RMSE = rmses[best_pvalue_idx])
```

4.3 Support Functions

4.3.1 fname00 Function

4.3.1.1 Usage

4.3.1.2 Parameters

::: { #func.fname00.params.param1 .sidebar } - **param1:** description;

- **param2:** description.

4.3.1.3 Details



Note

4.3.1.4 Value

4.3.1.5 Source Code

The source code of the fname00 function is shown below:



The source code of the fname00 is also available in the Some.Section section of the Somefunctions.R script.

4.3.2 fname01 Function

4.3.2.1 Signature

4.3.2.2 Parameters

- **param1:** description; - **param2:** description.

4.3.2.3 Details



Note

4.3.2.4 Value

4.3.2.5 Source Code

The source code of the function function is shown below:



The source code of the `fname0` function is also available in the `Regularization` section of the `UM-effect.functions.R` script.

4.3.3 fname02 Function

4.3.3.1 Signature

4.3.3.2 Parameters

- `param1`: description; - `param2`: description.

4.3.3.3 Details



Note

4.3.3.4 Value

4.3.3.5 Source Code

The source code of the function function is shown below:



The source code of the `fname02` function is also available in the `Regularization` section of the `UM-effect.functions.R` script.

4.3.4 fname0 Function

4.3.4.1 Signature

4.3.4.2 Parameters

- **param1**: description; - **param2**: description.

4.3.4.3 Details



Note

4.3.4.4 Value

4.3.4.5 Source Code

The source code of the function function is shown below:



The source code of the `function` is also available in the [Regularization](#) section of the [UM-effect.functions.R](#) script.

4.3.5 fname03 Function

4.3.5.1 Signature

4.3.5.2 Parameters

- **param1**: description; - **param2**: description.

4.3.5.3 Details



Note

4.3.5.4 Value

4.3.5.5 Source Code

The source code of the function function is shown below:



The source code of the function is also available in the Regularization section of the UM-effect.functions.R script.

4.3.6 fname1 Function

4.3.6.1 Signature

4.3.6.2 Parameters

- **param1**: description; - **param2**: description.

4.3.6.3 Details



Note

4.3.6.4 Value

4.3.6.5 Source Code

The source code of the function function is shown below:



The source code of the function is also available in the Regularization section of the UM-effect.functions.R script.

4.3.7 fname2 Function

4.3.7.1 Signature

4.3.7.2 Parameters

- **param1**: description; - **param2**: description.

4.3.7.3 Details



Note

4.3.7.4 Value

4.3.7.5 Source Code

The source code of the function function is shown below:



The source code of the `function` is also available in the Regularization section of the UM-effect.functions.R script.

4.3.8 fname3 Function

4.3.8.1 Signature

4.3.8.2 Parameters

- **param1**: description; - **param2**: description.

4.3.8.3 Details



Note

4.3.8.4 Value

4.3.8.5 Source Code

The source code of the function function is shown below:



The source code of the `function` is also available in the Regularization section of the UM-effect.functions.R script.

4.3.9 fname4 Function

4.3.9.1 Signature

4.3.9.2 Parameters

- **param1**: description; - **param2**: description.

4.3.9.3 Details



Note

4.3.9.4 Value

4.3.9.5 Source Code

The source code of the function function is shown below:



The source code of the `fname4` function is also available in the Regularization section of the `UM-effect.functions.R` script.

4.3.10 fname5 Function

4.3.10.1 Signature

4.3.10.2 Parameters

- **param1**: description; - **param2**: description.

4.3.10.3 Details



Note

4.3.10.4 Value

4.3.10.5 Source Code

The source code of the function function is shown below:



The source code of the function is also available in the Regularization section of the UM-effect.functions.R script.

4.3.11 fname6 Function

4.3.11.1 Signature

4.3.11.2 Parameters

- **param1**: description; - **param2**: description.

4.3.11.3 Details



Note

4.3.11.4 Value

4.3.11.5 Source Code

The source code of the function function is shown below:



The source code of the function is also available in the Regularization section of the UM-effect.functions.R script.

4.3.12 fname7 Function

4.3.12.1 Signature

4.3.12.2 Parameters

- **param1**: description; - **param2**: description.

4.3.12.3 Details



Note

4.3.12.4 Value

4.3.12.5 Source Code

The source code of the function function is shown below:



The source code of the `function` is also available in the Regularization section of the `UM-effect.functions.R` script.

4.3.13 fname8 Function

4.3.13.1 Signature

4.3.13.2 Parameters

- **param1**: description; - **param2**: description.

4.3.13.3 Details



Note

4.3.13.4 Value

4.3.13.5 Source Code

The source code of the function function is shown below:



The source code of the is also available in the Regularization section of the UM-effect.functions.R script.

References

- [1] Rafael A. Irizarry. *Introduction to Data Science, Part II, Chapter 23: Regularization, Section 23.2: Loss function. Statistics and Prediction Algorithms Through Case Studies*. Dec. 27, 2024. URL: <https://rafaelab.dfci.harvard.edu/dsbook-part-2/highdim/regularization.html#sec-netflix-loss-function> (visited on 02/18/2025) (cit. on p. 2).
- [2] Robert M. Bell Andreas Toscher Michael Jahrer. *The BigChaos Solution to the Netflix Grand Prize. commendo research & consulting*. Sept. 5, 2009. URL: https://www.asc.ohio-state.edu/statistics/statgen/joul_au2009/BigChaos.pdf (visited on 02/18/2025) (cit. on pp. 2, 31).
- [3] Azamat Kurbanayev. *edX Data Science: Capstone, MovieLens Datasets. Package: edx.capstone.movielens.data*. Version 0.0.0.9000. Feb. 5, 2025. URL: <https://github.com/AzKurban-edX-DS/edx.capstone.movielens.data> (visited on 02/05/2025) (cit. on p. 3).
- [4] Rafael A. Irizarry. *Introduction to Data Science, Part II. Statistics and Prediction Algorithms Through Case Studies*. Dec. 27, 2024. URL: <https://rafaelab.dfci.harvard.edu/dsbook-part-2/> (visited on 02/18/2025) (cit. on p. 4).
- [5] Rafael A. Irizarry. *Introduction to Data Science, Part II, Chapter 23: Regularization, Section 23.1.1: MovieLens data. Statistics and Prediction Algorithms Through Case Studies*. Dec. 27, 2024. URL: <https://rafaelab.dfci.harvard.edu/dsbook-part-2/highdim/regularization.html#movielens-data> (visited on 02/18/2025) (cit. on pp. 5, 17, 19).
- [6] Amir Motefaker. *Movie Recommendation System using R - BEST*. Version 284. July 18, 2024. URL: <https://www.kaggle.com/code/amirmotefaker/movie-recommendation-system-using-r-best/notebook> (visited on 02/18/2025) (cit. on pp. 7–12, 47, 49).
- [7] Azamat Kurbanayev. *edX Data Science: Capstone-MovieLens Project. A movie recommendation system using the MovieLens dataset*. Version 1.0.0.0. May 5, 2025. URL: <https://github.com/AzKurban-edX-DS/Capstone-MovieLens/tree/main> (visited on 05/05/2025) (cit. on p. 15).
- [8] Rafael A. Irizarry. *Introduction to Data Science, Part II, Chapter 23: Regularization, Section 23.3: A first model. Statistics and Prediction Algorithms Through Case Studies*. Dec. 27, 2024. URL: <https://rafaelab.dfci.harvard.edu/dsbook-part-2/highdim/regularization.html#a-first-model> (visited on 02/18/2025) (cit. on p. 31).
- [9] Rafael A. Irizarry. *Introduction to Data Science, Part II, Chapter 23: Regularization, Section 23.4: User effects. Statistics and Prediction Algorithms Through Case Studies*. Dec. 27, 2024. URL: <https://rafaelab.dfci.harvard.edu/dsbook-part-2/highdim/regularization.html#user-effects> (visited on 02/18/2025) (cit. on p. 32).
- [10] Robert Bell Yehuda Koren Yahoo Research and Chris Volinsky. *Matrix Factorization Techniques for Recommender Systems*. Aug. 1, 2009. URL: [https://datajobs.com/data-science-repo/Recommender-Systems-\[Netflix\].pdf](https://datajobs.com/data-science-repo/Recommender-Systems-[Netflix].pdf) (visited on 02/18/2025) (cit. on p. 32).
- [11] Rafael A. Irizarry. *Introduction to Data Science, Part II, Chapter 23: Regularization, Section 23.5: Movie effects. Statistics and Prediction Algorithms Through Case Studies*. Dec. 27, 2024. URL: <https://rafaelab.dfci.harvard.edu/dsbook-part-2/highdim/regularization.html#movie-effects> (visited on 02/18/2025) (cit. on p. 40).

- [12] Rafael A. Irizarry. *Introduction to Data Science, Part II, Chapter 23: Regularization, Section 23.6: Penalized least squares. Statistics and Prediction Algorithms Through Case Studies*. Dec. 27, 2024. URL: <https://rafalab.dfci.harvard.edu/dsbook-part-2/highdim/regularization.html#penalized-least-squares> (visited on 02/18/2025) (cit. on pp. 41, 57).
- [13] Rafael A. Irizarry. *Introduction to Data Science, Part II, Chapter 23: Regularization, Section 23.7: Exercises. Statistics and Prediction Algorithms Through Case Studies*. Dec. 27, 2024. URL: <https://rafalab.dfci.harvard.edu/dsbook-part-2/highdim/regularization.html#exercises> (visited on 02/18/2025) (cit. on pp. 46, 48).
- [14] Francesco Ricci. *Recommender Systems Handbook*. Ed. by Paul B. Kantor Lior Rokach Bracha Shapira. Springer, New York, 2011. ISBN: ISBN 978-0-387-85819-7. DOI: [10.1007/978-0-387-85820-3](https://doi.org/10.1007/978-0-387-85820-3). URL: https://github.com/vwang0/recommender_system/blob/master/Recommender%20Systems%20Handbook.pdf (cit. on p. 48).