StudyCase: Netflix RS

Diego Lesmes

10/11/2020

```
library(tidyverse)
library(caret)
library(data.table)
library(knitr)
```

EDA MovieLens Dataset

First is developed the exploratory data analysis to understand the data setup and the data structure is available from here.

So lets go to start:

Data Available

```
edx <- get("edx")
str(edx)
```

```
## Classes 'data.table' and 'data.frame':   9000055 obs. of  6 variables:
##  $ userId   : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ movieId  : num  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 83
##  $ title    : chr  "Boomerang (1992)" "Net, The (1995)" "Outbreak (1995)" "Stargate (1994)" ...
##  $ genres   : chr  "Comedy|Romance" "Action|Crime|Thriller" "Action|Drama|Sci-Fi|Thriller" "Action|Ac
##  - attr(*, ".internal.selfref")=<externalptr>
```

```
summary(edx) %>% knitr::kable()
```

userId	movieId	rating	timestamp	title	genres
Min. : 1	Min. : 1	Min. :0.500	Min. :7.897e+08	Length:9000055	Length:9000055
1st Qu.:18124	1st Qu.: 648	1st Qu.:3.000	1st Qu.:9.468e+08	Class :character	Class :character
Median :35738	Median : 1834	Median :4.000	Median :1.035e+09	Mode :character	Mode :character
Mean :35870	Mean : 4122	Mean :3.512	Mean :1.033e+09	NA	NA

userId	movieId	rating	timestamp	title	genres
3rd Qu.:53607	3rd Qu.: 3626	3rd Qu.:4.000	3rd Qu.:1.127e+09	NA	NA
Max. :71567	Max. :65133	Max. :5.000	Max. :1.231e+09	NA	NA

```
dim(edx)
```

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

Variables

Ratings

```
edx %>% group_by(rating) %>%
  summarise(N_Score = n()) %>%
  knitr::kable()
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

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

```
hist(edx$rating,col = "light yellow",
     main = "Ratings Predictor",
     xlab = "Rating")
```

Movies

```
length(unique(edx$movieId))
```

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

Users

```
length(unique(edx$userId))
```

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

Genres

Number of ratings are in each of the following genres

- Drama

```
edx %>% filter(str_detect(genres,"Drama")) %>%
  nrow()
```

```
## [1] 3910127
```

- Comedy

```
edx %>% filter(str_detect(genres,"Comedy")) %>%
  nrow()
```

```
## [1] 3540930
```

- Thriller

```
edx %>% filter(str_detect(genres,"Thriller")) %>%
  nrow()
```

```
## [1] 2325899
```

- Romance

```
edx %>% filter(str_detect(genres,"Romance")) %>%
  nrow()
```

```
## [1] 1712100
```

Title

The movie with the greatest number of ratings is

```
edx %>% group_by(title) %>%
  summarise("N_Ratings" = n()) %>%
  arrange(N_Ratings) %>%
  tail()
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 6 x 2
##   title                              N_Ratings
##   <chr>                                  <int>
## 1 Braveheart (1995)                      26212
## 2 Shawshank Redemption, The (1994)       28015
## 3 Jurassic Park (1993)                   29360
## 4 Silence of the Lambs, The (1991)       30382
## 5 Forrest Gump (1994)                    31079
## 6 Pulp Fiction (1994)                    31362
```