CAPSTONE-PH125.9x

Laxmansingh Rajput

5/31/2019

OBJECTIVES

Creating a movie recommendation system using the MovieLens dataset. Train a machine learning algorithm using the inputs in one subset to predict movie ratings in the validation set.

This project is being completed in R version 3.6.0 (*Planting of a Tree*)

The submission for the MovieLens project will be three files:

- A report in the form of an Rmd file
- A report in the form of a PDF document knit from your Rmd file
- An Rmd/R script file that generates the predicted movie ratings and calculates RMSE

The movie rating predictions will be compared to the true ratings in the validation set using RMSE and the report outputs the RMSE.

EXECUTIVE SUMMARY

```
1. Dataset
   a. MovieLens dataset (downloaded as R Version is 3.6.0)
      from the following google drive
      <https://drive.google.com/drive/folders/1IZcBBX0OmL9wu9AdzMBFUG8GoPbGQ38D>
2. Goals of the project
   a. Predict the ratings for the validation dataset
   b. Compute the RMSE
3. Key Steps
   a. Compute the average of the movie ratings
   b. Calculate the requlalrized parameter (lambda) using cross-validation
   c. Get the lambda that gives the least RMSE
   d. Use the lambda to predict the ratings
   e. Compute the RMSE on the predict ratings vs the validation test
```

METHOD/ANALYSIS SECTION

- Explore the data
- Explore the rating information
- Explore the top rated movies
- Model selection - use the regularization method on the rating based on movieID and userID biases

FUNCTIONS TO BE USED

- RMSE <- function(true_ratings, predicted_ratings) sqrt(mean((true_ratings - predicted_ratings)^2))
- predict_rating <- compute the movie and user biases, and return the predicted ratings based on the regularization parameter.

EXPLORE THE DATASET

We can examine the edx data by using R's `glimpse` command:

```
Observations: 9,000,055
Variables: 6
$ userId    <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1...
$ movieId   <dbl> 122, 185, 292, 316, 329, 355, 356, 362, 364, 370, 37...
$ rating    <dbl> 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5...
$ timestamp <int> 838985046, 838983525, 838983421, 838983392, 83898339...
$ title     <chr> "Boomerang (1992)", "Net, The (1995)", "Outbreak (19...
$ genres    <chr> "Comedy|Romance", "Action|Crime|Thriller", "Action|D...
```

Here we see that we have a data frame, accessible at `edx`:

```
  userId movieId rating timestamp                         title
1      1     122      5 838985046              Boomerang (1992)
2      1     185      5 838983525               Net, The (1995)
4      1     292      5 838983421               Outbreak (1995)
5      1     316      5 838983392               Stargate (1994)
6      1     329      5 838983392 Star Trek: Generations (1994)
7      1     355      5 838984474       Flintstones, The (1994)
                         genres
1                Comedy|Romance
2          Action|Crime|Thriller
4  Action|Drama|Sci-Fi|Thriller
5       Action|Adventure|Sci-Fi
6 Action|Adventure|Drama|Sci-Fi
7       Children|Comedy|Fantasy
```

Understanding Variables

1. The list of the variables.

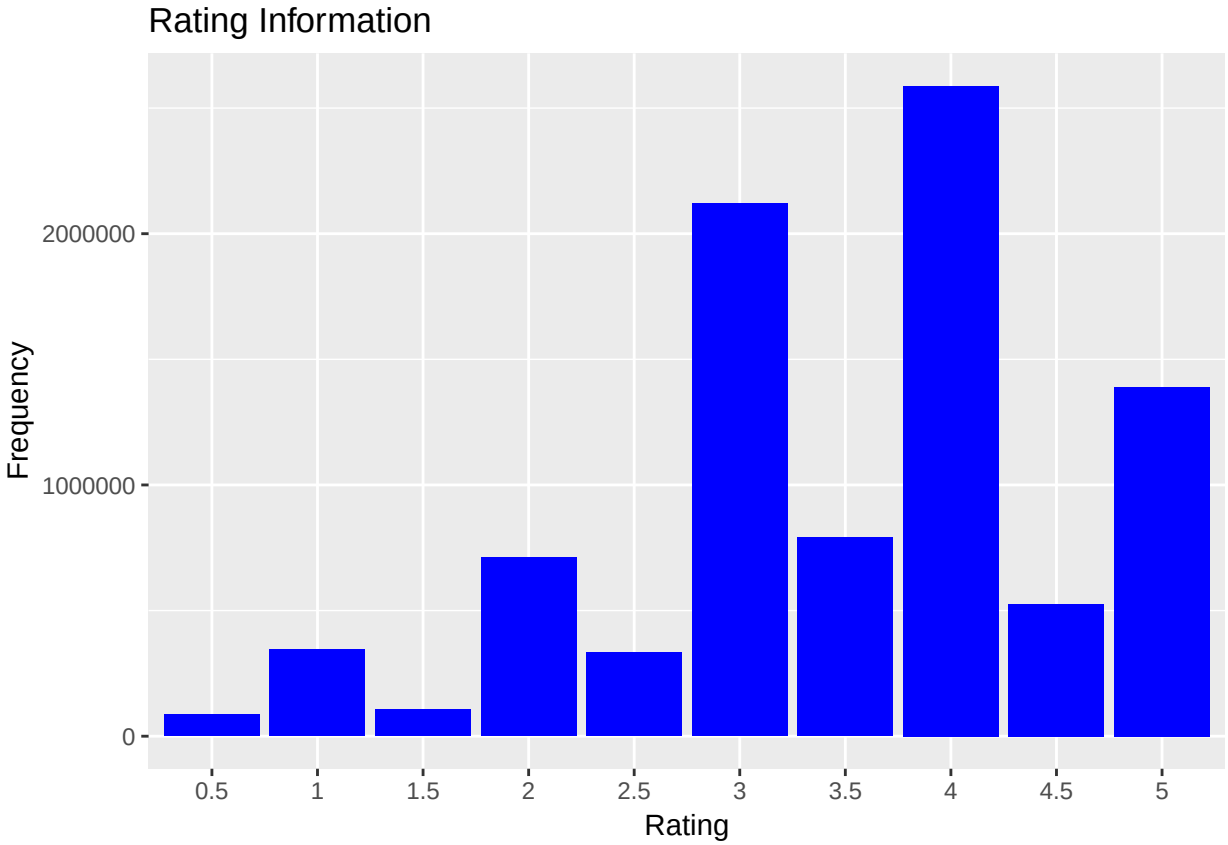
Table 1: Variable Names

userId	movieId	rating	timestamp	title	genres

2. Other information regarding the dataset - Observations, variables, unique values, and mean rating

Table 2: Variable Information

DESCRIPTION	VALUES
Observations	9000055
Variables	6
Unique User IDs	69878
Unique Movie IDs	10677
Unique Ratings	10
Mean Rating	3.512465



3. Rating information

Table 3: Rating Information

Rating	0.5	1	1.5	2	2.5	3	3.5	4	4.5	5
Frequency	85374	345679	106426	711422	333010	2121240	791624	2588430	526736	1390114

Note:
Overall Summary of the ratings.
[1] Rating ranges from 0.5 to 5 in 0.5 steps

Table 4: Rating in reverse order

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

4. Five most rated movies

Table 5: Top 5 rated movies

Title	Frequency
Pulp Fiction (1994)	31362
Forrest Gump (1994)	31079
Silence of the Lambs, The (1991)	30382
Jurassic Park (1993)	29360
Shawshank Redemption, The (1994)	28015

5. edx dataset

Table 6: edx Dataset (Snapshot)

	userId	movieId	rating	title	genres
1	1	122	5	Boomerang (1992)	Comedy|Romance
2	1	185	5	Net, The (1995)	Action|Crime|Thriller
4	1	292	5	Outbreak (1995)	Action|Drama|Sci-Fi|Thriller
5	1	316	5	Stargate (1994)	Action|Adventure|Sci-Fi
6	1	329	5	Star Trek: Generations (1994)	Action|Adventure|Drama|Sci-Fi
7	1	355	5	Flintstones, The (1994)	Children|Comedy|Fantasy
8	1	356	5	Forrest Gump (1994)	Comedy|Drama|Romance|War
9	1	362	5	Jungle Book, The (1994)	Adventure|Children|Romance
10	1	364	5	Lion King, The (1994)	Adventure|Animation|Children|Drama|Musical
11	1	370	5	Naked Gun 33 1/3: The Final Insult (1994)	Action|Comedy
12	1	377	5	Speed (1994)	Action|Romance|Thriller
13	1	420	5	Beverly Hills Cop III (1994)	Action|Comedy|Crime|Thriller
14	1	466	5	Hot Shots! Part Deux (1993)	Action|Comedy|War
16	1	520	5	Robin Hood: Men in Tights (1993)	Comedy
17	1	539	5	Sleepless in Seattle (1993)	Comedy|Drama|Romance
19	1	588	5	Aladdin (1992)	Adventure|Animation|Children|Comedy|Musical
20	1	589	5	Terminator 2: Judgment Day (1991)	Action|Sci-Fi
21	1	594	5	Snow White and the Seven Dwarfs (1937)	Animation|Children|Drama|Fantasy|Musical
22	1	616	5	Aristocats, The (1970)	Animation|Children
23	2	110	5	Braveheart (1995)	Action|Drama|War

6. validation dataset

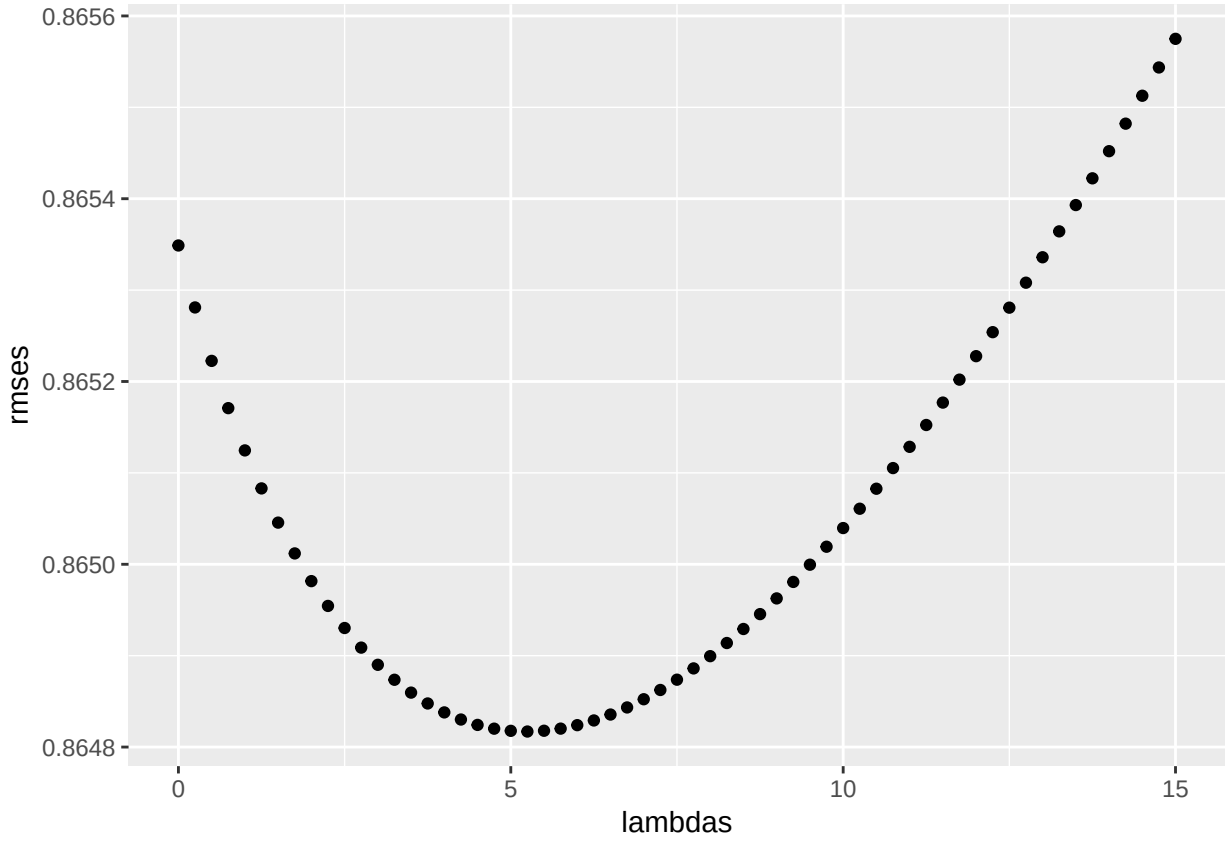
Table 7: validation Dataset (Snapshot)

userId	movieId	rating	title	genres
1	231	5.0	Dumb & Dumber (1994)	Comedy
1	480	5.0	Jurassic Park (1993)	Action|Adventure|Sci-Fi|Thriller
1	586	5.0	Home Alone (1990)	Children|Comedy
2	151	3.0	Rob Roy (1995)	Action|Drama|Romance|War
2	858	2.0	Godfather, The (1972)	Crime|Drama
2	1544	3.0	Lost World: Jurassic Park, The (Jurassic Park 2) (1997)	Action|Adventure|Horror|Sci-Fi|Thriller
3	590	3.5	Dances with Wolves (1990)	Adventure|Drama|Western
3	4995	4.5	Beautiful Mind, A (2001)	Drama|Mystery|Romance
4	34	5.0	Babe (1995)	Children|Comedy|Drama|Fantasy
4	432	3.0	City Slickers II: The Legend of Curly's Gold (1994)	Adventure|Comedy|Western
4	434	3.0	Cliffhanger (1993)	Action|Adventure|Thriller
5	85	3.0	Angels and Insects (1995)	Drama|Romance
5	171	3.0	Jeffrey (1995)	Comedy|Drama
5	232	3.0	Eat Drink Man Woman (Yin shi nan nu) (1994)	Comedy|Drama|Romance
5	242	3.0	Farinelli: il castrato (1994)	Drama|Musical
5	306	3.0	Three Colors: Red (Trois couleurs: Rouge) (1994)	Drama
5	345	3.0	Adventures of Priscilla, Queen of the Desert, The (1994)	Comedy|Drama
5	592	3.0	Batman (1989)	Action|Crime|Sci-Fi|Thriller
5	593	4.0	Silence of the Lambs, The (1991)	Crime|Horror|Thriller
5	920	5.0	Gone with the Wind (1939)	Drama|Romance|War

CROSS-VALIDATION TO PICK A REGULARIZATION PARAMETER (LAMBDA)

MOVIE + USER EFFECT

```
Use sapply function to return a vector and then retrieve the
  parameter that gives the minimum RMSE
     1. Calculate the regularized bias for each movies
     2. Calculate the regularized bias for each user
     3. Predict ratings based on both regularized movies & user biases
     4. Compare against the validation (test) set and compute the RMSE
```



RESULTS SECTION

- Best parameter
- Regularization parameters and RMSE

```
[1] "Best regularization parameter:  5.25"
```

Regularization parameters and RMSE

	LAMBDAS	RMSE
1	0	0.865348824577316
2	0.25	0.865281027927129
3	0.5	0.86522255159723
4	0.75	0.86517084725625
5	1	0.865124604840233
6	1.25	0.865083030819814
7	1.5	0.865045585712057
8	1.75	0.865011871216156
9	2	0.86498157524896
10	2.25	0.864954442465168
11	2.5	0.864930257176986
12	2.75	0.864908832811132
13	3	0.86489000504715
14	3.25	0.864873627146169
15	3.5	0.864859566645536
16	3.75	0.864847702940283
17	4	0.864837925461094
18	4.25	0.864830132266144
19	4.5	0.8648242289282
20	4.75	0.864820127637599
21	5	0.864817746466669
22	5.25	0.864817008757297
23	5.5	0.864817842604197
24	5.75	0.864820180413763
25	6	0.864823958523533
26	6.25	0.86482911687093
27	6.5	0.864835598702562
28	6.75	0.864843350317273
29	7	0.864852320837594
30	7.25	0.864862462005291
31	7.5	0.864873727997555

	LAMBDAS	RMSE
32	7.75	0.864886075260981
33	8	0.864899462361034
34	8.25	0.864913849845025
35	8.5	0.864929200116996
36	8.75	0.864945477323118
37	9	0.864962647246444
38	9.25	0.864980677210002
39	9.5	0.864999535987365
40	9.75	0.865019193719952
41	10	0.865039621840395
42	10.25	0.865060793001412
43	10.5	0.865082681009668
44	10.75	0.865105260764182
45	11	0.865128508198891
46	11.25	0.865152400229008
47	11.5	0.865176914700864
48	11.75	0.865202030344954
49	12	0.865227726731926
50	12.25	0.865253984231291
51	12.5	0.865280783972639
52	12.75	0.865308107809182
53	13	0.865335938283448
54	13.25	0.865364258594973
55	13.5	0.865393052569844
56	13.75	0.865422304631973
57	14	0.865451999775976
58	14.25	0.86548212354155
59	14.5	0.865512661989251
60	14.75	0.865543601677574
61	15	0.865574929641264

CONCLUSION SECTION

- Best RMSE
- Some predicted ratings

```
[1] "Best RMSE:  0.864817008757297"
```

PREDICTED RATING (SNAPSHOT)

Table 8: PREDICTED RATINGS (Snapshot)

userId	movieId	rating	title	predicted_ratings
1	231	5.0	Dumb & Dumber (1994)	4.250809
1	480	5.0	Jurassic Park (1993)	4.978994
1	586	5.0	Home Alone (1990)	4.371325
2	151	3.0	Rob Roy (1995)	3.349371
2	858	2.0	Godfather, The (1972)	4.234425
2	1544	3.0	Lost World: Jurassic Park, The (Jurassic Park 2) (1997)	2.765010
3	590	3.5	Dances with Wolves (1990)	3.969120
3	4995	4.5	Beautiful Mind, A (2001)	4.132423
4	34	5.0	Babe (1995)	4.272564
4	432	3.0	City Slickers II: The Legend of Curly's Gold (1994)	3.301982
4	434	3.0	Cliffhanger (1993)	3.642678
5	85	3.0	Angels and Insects (1995)	3.590682
5	171	3.0	Jeffrey (1995)	3.676414
5	232	3.0	Eat Drink Man Woman (Yin shi nan nu) (1994)	4.146335
5	242	3.0	Farinelli: il castrato (1994)	3.483259
5	306	3.0	Three Colors: Red (Trois couleurs: Rouge) (1994)	4.220861
5	345	3.0	Adventures of Priscilla, Queen of the Desert, The (1994)	3.750159
5	592	3.0	Batman (1989)	3.466363
5	593	4.0	Silence of the Lambs, The (1991)	4.284026
5	920	5.0	Gone with the Wind (1939)	3.875768

Information About The Current R Session

```
R version 3.6.0 (2019-04-26)
Platform: x86_64-w64-mingw32/x64 (64-bit)
Running under: Windows 10 x64 (build 14393)

Matrix products: default

Random number generation:
 RNG:     Mersenne-Twister
 Normal:  Inversion
 Sample:  Rounding

locale:
[1] LC_COLLATE=English_United States.1252
[2] LC_CTYPE=English_United States.1252
[3] LC_MONETARY=English_United States.1252
[4] LC_NUMERIC=C
[5] LC_TIME=English_United States.1252

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
 [1] gridExtra_2.3     googledrive_0.1.3 knitr_1.23
 [4] kableExtra_1.1.0  forcats_0.4.0     stringr_1.4.0
 [7] dplyr_0.8.1       purrr_0.3.2       readr_1.3.1
[10] tidyr_0.8.3       tibble_2.1.1      ggplot2_3.1.1
[13] tidyverse_1.2.1

loaded via a namespace (and not attached):
 [1] tidyselect_0.2.5  xfun_0.7          haven_2.1.0
 [4] lattice_0.20-38   colorspace_1.4-1  generics_0.0.2
 [7] htmltools_0.3.6   viridisLite_0.3.0 yaml_2.2.0
[10] rlang_0.3.4       pillar_1.4.0      glue_1.3.1
[13] withr_2.1.2       modelr_0.1.4      readxl_1.3.1
[16] plyr_1.8.4        munsell_0.5.0     gtable_0.3.0
[19] cellranger_1.1.0  rvest_0.3.4       codetools_0.2-16
[22] evaluate_0.13     broom_0.5.2       Rcpp_1.0.1
[25] scales_1.0.0      backports_1.1.4   webshot_0.5.1
[28] jsonlite_1.6      hms_0.4.2         digest_0.6.19
[31] stringi_1.4.3     grid_3.6.0        cli_1.1.0
[34] tools_3.6.0       magrittr_1.5      lazyeval_0.2.2
[37] crayon_1.3.4      pkgconfig_2.0.2   xml2_1.2.0
[40] lubridate_1.7.4   assertthat_0.2.1  rmarkdown_1.12
[43] httr_1.4.0        rstudioapi_0.10   R6_2.4.0
[46] nlme_3.1-139      compiler_3.6.0
```