MovieLens Project Final Report

Khaliun Bat-Ochir

May 4th, 2021

Contents

1 Overview

This is the MovieLens Project Final Report for Capstone Course from Data Science Professional Certificate Program offered by HarvardX and online initiative by Harvard University prepared by Khaliun Bat-Ochir.

Final Report includes methodology and results of the project including RMSE calculation. Also, this report includes project environment description used for building and running the code.

1.1 Project Goal

This is a Machine Learning project. Goal of the project consists in analysing **MovieLens 10M** and training data with chosen Machine Learning algorithm and reaching RMSE below 0.86490.

1.2 Project Data

Project uses **MovieLens 10M** data set. The original code from provided by the instruction for Movielens Recommendation system project for Capstone Course downloads original data and as a *Zip File*[1]. Code then processes separate two files named **ratings** and **movies** included in the zip file that are joined together by "movieId" field.

1.3 Project Files

Project files will be uploaded for review and grading by peers to Edx course section. Uploaded Files Include:

- RMarkdown Report File - "MovieLensAnalysisReport.Rmd"
- R Script File - for convenience, R scripts include + MovieLensAnalysisScripts.R - which contains data preparation scripts + MovieLensAnalysisScripts_DataExploration.R - which contains analysis codes related to this report. (Please note that this file sources MovieLensAnalysisScripts.R for running). + and MovieLensAnalysisScripts_RMSE_edx_dataset.R" that includes codes for model fitting and RMSE calculation for final model. (Please note that this file sources MovieLensAnalysisScripts.R for running).
- PDF Report - "MovieLensAnalysisReport.pdf"

For convenience purposes, total runing time of each model fitting algorithm had been included in Result section of this report.

All the files have been also uploaded to *MovieLensCodes Github Repository*[2]

1.4 Project Environment

Codes for the project were built and tested using:

- R version 3.6 and

[1] http://files.grouplens.org/datasets/movielens/ml-10m.zip
[2] https://github.com/khaliunb/MovieLensCodes.git

- RStudio Version 1.3.1073
- Linux Ubuntu 20.04

2 Method

2.1 Original Code and Data Set Preparation

Original code divides downloaded data set into following two subsets:

- **edx** - equivalent of **training set** set that contains 80% of the complete **MovieLens10K** data set
- **validation** - equivalent of **test set** that contains 20% of the complete **MovieLens10K** data set

Furthermore, **edx** data set had been matched with **validation** data set and all recurring data had been removed from **validation** data set.

Finally, the code removes temporary data sets (dl, ratings, movies, test_index, temp, movielens, removed) that were used to prepare **edx** and **validation** data sets.

Summary of **edx** data set:

```
summary(edx%>%select(userId, movieId,rating, timestamp,title,genres))
```

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

Summary of **validation** data set:

```
##      userId         movieId          rating        timestamp
##  Min.   :    1   Min.   :    1   Min.   :0.500   Min.   :7.897e+08
##  1st Qu.:18096   1st Qu.:  648   1st Qu.:3.000   1st Qu.:9.467e+08
##  Median :35768   Median : 1827   Median :4.000   Median :1.035e+09
##  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
##     title              genres
##  Length:999999      Length:999999
```

```
##  Class :character   Class :character
##  Mode  :character   Mode  :character
##
##
##
```

Note: Only **edx** data set have been used for data exploration, model fitting and training the data. **validation** data set will be used for final results and RMSE calculation.

2.2 Data Exploration and Sampling method

2.2.1 Linear model analysis data

Linear model analysis used for Machine Learning course will be applied for complete **edx** data set and results will be used for final model.

Regularized movie and user effect model have also been attempted on sample of 10'000.

However, the results were inconclusive as the sample data did not represent the completed **edx** data set. Therefore final *lambda* value for the model and analysis for the linear model have been carried out using complete **edx** data set.

train_set and **test_set** partitioned from complete **edx** data set had been used.

- **train_set** - contains 80% of the complete **edx** data set
- **test_set** - contains 20% of the complete **edx** data set

Furthermore, **train_set** data set had been matched with **test_set** data set and all recurring data had been removed from **validation** data set.

Summary of **train_set**:

```
##      userId         movieId          rating        timestamp
##  Min.   :    1   Min.   :    1   Min.   :0.500   Min.   :7.897e+08
##  1st Qu.:18111   1st Qu.:  648   1st Qu.:3.000   1st Qu.:9.467e+08
##  Median :35736   Median : 1834   Median :4.000   Median :1.036e+09
##  Mean   :35867   Mean   : 4123   Mean   :3.513   Mean   :1.033e+09
##  3rd Qu.:53609   3rd Qu.: 3624   3rd Qu.:4.000   3rd Qu.:1.127e+09
##  Max.   :71567   Max.   :65133   Max.   :5.000   Max.   :1.231e+09
##     title              genres
##  Length:7200043     Length:7200043
##  Class :character   Class :character
##  Mode  :character   Mode  :character
##
##
##
```

Summary of **test_set**:

```
##      userId         movieId          rating        timestamp
##  Min.   :    1   Min.   :    1   Min.   :0.500   Min.   :8.229e+08
##  1st Qu.:18167   1st Qu.:  648   1st Qu.:3.000   1st Qu.:9.468e+08
##  Median :35739   Median : 1834   Median :4.000   Median :1.035e+09
##  Mean   :35880   Mean   : 4116   Mean   :3.512   Mean   :1.033e+09
##  3rd Qu.:53596   3rd Qu.: 3633   3rd Qu.:4.000   3rd Qu.:1.127e+09
##  Max.   :71567   Max.   :65133   Max.   :5.000   Max.   :1.231e+09
##     title              genres
##  Length:1799968     Length:1799968
##  Class :character   Class :character
##  Mode  :character   Mode  :character
##
##
##
```

2.3 Failures and Insights gained

Kmeans clustering have been tried for grouping movies into relevant groups as an attempt to replicate PCA analysis and incorporated the group ids to **edx** data set. However, the final RMSE for movie group model is 0.99 and does not go down for any values of k. Therefore the approach have been dropped. Code performing this analysis have been commented out in file * MovieAnalysisScripts_DataExploration.R *

2.4 Model fitting and RMSE approach

Regularized movie and user effect model have be used for final model. For this, we will be using full cross-validation with **train_set** and **test_set** data sets for *lambda* parameter tuning. From there, minimum value of RMSE achieving *lambda* will be used final model training on **edx** data set.

For Final RMSE **validation** set will be used.

3 Results

Note: Results section contains plots and summary results present in code file *MovieAnalysisScripts_DataExploration.R* You can find relevant comments as a description for each result in this file.

3.1 Linear model performance

We are getting a peek into full **edx** data set. For this, we are summarizing number of users and movies present in the data set.

```
##   n_users n_movies
## 1   69878    10677
```

Now, let us list top 5 most rated movies in movielens data

```
keep
```

```
## [1] 296 318 356 480 593
```

Top 5 most rated movies' ratings in movielens data and transposes the title and rating columns by value and lists the results for userId column.

```
tab %>% knitr::kable()
```

userId	Forrest Gump (1994)	Jurassic Park (1993)	Pulp Fiction (1994)	Silence of the Lambs, The (1991)
1	5	NA	NA	NA
4	NA	5	NA	NA
7	NA	NA	NA	3
8	NA	3	NA	4
10	3	NA	2	3
11	NA	4	3	NA

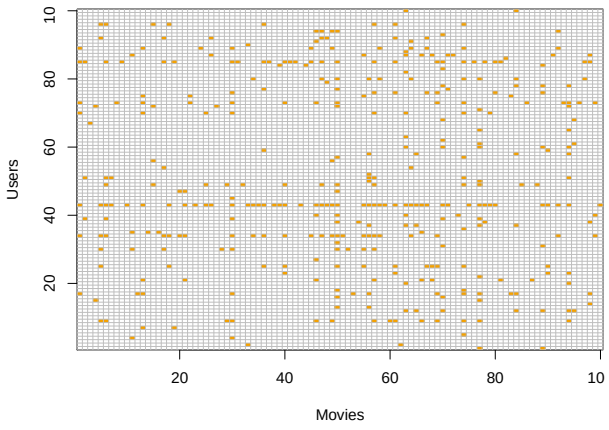
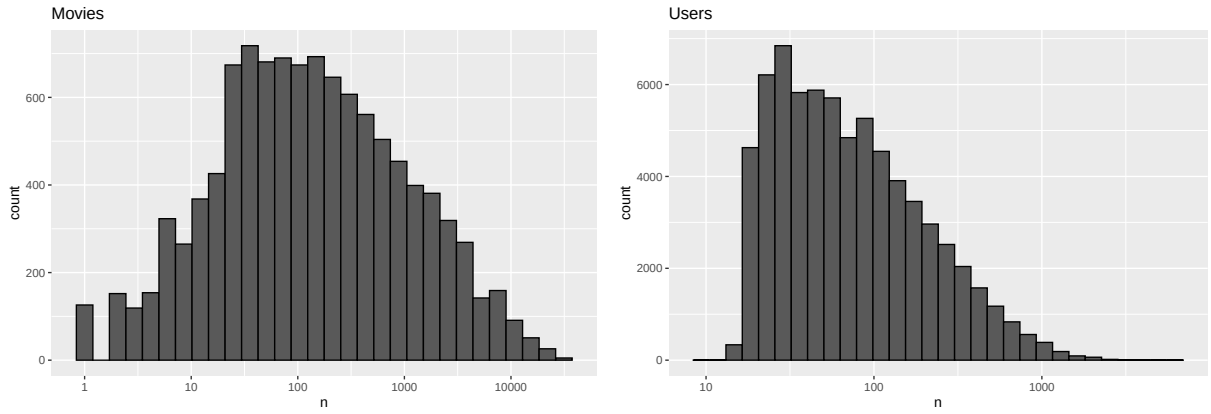


Figure 1: Imaging of the movies and users rating in 100 pixels

We are showing the density plots for distribution of Movies and Users ratings.



(a) Movies Ratings Density (b) Users Ratings Density

Figure 2: Density Plots for Rating: (a) Movies (b) Users

As we can see, the (b) Users Ratings Density plot differs from the Machine Learning course analysis results. Users are no longer rating mostly 1 time. This can be attributed to the fact that we are using more complete data set *MovieLens10M*.

3.1.1 lambda value

We will remember that *lambda* is a tuning parameter. Therefore we will use cross-validation to choose it and apply the final minimum RMSE value of *lambda* for final training.

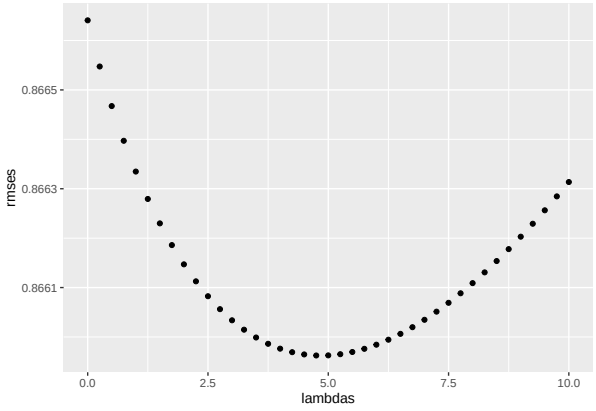


Figure 3: Qplot: Lambda against RMSE

Analysis results for the Regularized Movie + User Effect model shows that the optimal value of lambda is as follows:

```
## [1] 4.75
```

This value we will incorporate in the actual training of the linear model.

3.1.2 Regularized Movie + User Effect model RMSE

Result for minimum RMSE gained from tuning lambda for Regularized Movie + User Effect is as follows:

method	RMSE
Regularized Movie + User Effect Model	0.8659626

3.2 Final results, steps and timing of training on edx data set

We are using the *lambda* value gained using cross validation. We are incorporating the *mu*, *b_i*, *b_u*, *pred* features back into **edx** and **validation** data sets in file **MovieLensAnalysisScripts.R**. We will be only using *pred* feature for prediction and training Data preparation script runs approximately 2 minutes.

Then, we are fitting the model using *lm()* in file **MovieLensAnalysisScripts_RMSE_edx_dataset.R**.

Training of complete **edx** set was performed. Final RMSE is as follows:

method	FINAL_RMSE
Regularized Movie + User Effect Model	0.8648617

Total time of execution took a second.

log	DURATION
Data source prep script run time	132.39938 secs
Edx Data set training script run time	1.17256 secs

However, we should note that running time depends on the environment.

4 Conclusion

For final result, Regularized movie and user effect model have been used. Final RMSE is *0.8648617*. Achieved project goal of **RMSE < 0.86490**.

Contact Information

If you have any questions regarding the project, please feel free to contact me at any of my emails: *khaliun83@yahoo.com*[3], *khaliun@spoon.mn*[4]; or feel free to visit my *Linkedin Profile*[5]

References

Books

- Irizarry (2021)
- Xie, Dervieux, & Riederer (2020)

Articles

- Boehmke (2021)

Manuals

- R Core Team (2020)
- Wickham et al. (2021)
- Wickham et al. (2020)
- Wickham (2019)
- Datacamp team (2020)

Boehmke, B. (2021). UC Business Analytics R Programming Guide: K-Means Cluster Analysis. Retrieved from https://uc-r.github.io/kmeans_clustering

Datacamp team. (2020). *Predict.lm: Predict Method for Linear Model Fits.* Datacamp. Retrieved from https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/predict.lm

Irizarry, R. A. (2021). *Introduction to Data Science: Data Analysis and Prediction Algorithms with R.* CRC Press. Retrieved from https://rafalab.github.io/dsbook/

R Core Team. (2020). *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Wickham, H. (2019). *Tidyverse: Easily Install and Load the Tidyverse.* Retrieved from https://CRAN.R-project.org/package=tidyverse

Wickham, H., Chang, W., Henry, L., Pedersen, T. L., Takahashi, K., Wilke, C., Woo, K., Yutani, H. & Dunnington, D. (2020). *Ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics.* Retrieved from https://CRAN.R-project.org/package=ggplot2

[3] mailto:khaliun83@yahoo.com
[4] mailto:khaliun@spoon.mn
[5] https://www.linkedin.com/in/khaliun-bat-ochir-334925b4/

Wickham, H., François, R., Henry, L. & Müller, K. (2021). *Dplyr: A Grammar of Data Manipulation.* Retrieved from https://CRAN.R-project.org/package=dplyr

Xie, Y., Dervieux, C. & Riederer, E. (2020). *The R Series: RMarkdown Cookbook.* CRC Press Tayler; Francis Group A Chapman And Hall Book. Retrieved from https://bookdown.org/yihui/rmarkdown-cookbook/