Final project in Data Science at Harvard University
MovieLens

Jan Thomsen

01/08/2021

Contents

Abstract: "This is the final assignment for the Harvard Data Science Professional certificate Programme with Professor of Biostatistics Rafael Irizarry from Harvard University.

It is the 9th and last course in the Data Science series offered by Harvard University:

- **1. R basics**
- **2. Visualization**
- **3. Probability**
- **4. Inference and modeling**
- **5. Productivity tools**
- **6. Wrangling**
- **7. Linear regression**
- **8. Machine learning**
- **9. Capstone**

In this capstone project, we given the dataset and instructions we have to clean, analyze and modeling it and show our Data Science knowledge."

Chapter 1

Executive Summary

For achieving the task of analysing the dataset i have used knowledge obtained in the 8 courses.

Instructions from HarvardX this report should include:

- **introduction/overview/executive** summary section that describes the dataset and summarizes the goal of the project and key steps that were performed
- **methods/analysis** section that *explains* the process and techniques used:
- **data cleaning**
- **data exploration**
- **visualization**
- **results** section that presents the modeling results and discusses the model performance a conclusion section that gives a brief summary of the report, its limitations and future work

Chapter 2

Exploratory Data Analysis

MovieLens data sets were collected by the GroupLens Research Project at the University of Minnesota.

The data was collected through the MovieLens web site (movielens.umn.edu) during the seven-month period from September 19th, 1997 through April 22nd, 1998. This data has been cleaned up - users, who had less than 20 ratings or did not have complete demographic information were removed from this data set.

These data were created by 610 users between March 29, 1996 and September 24, 2018. This dataset was generated on September 26, 2018.

Users were selected at random for inclusion. All selected users had rated at least 20 movies. No demographic information is included. Each user is represented by an id, and no other information is provided.

This is a development dataset. As such, it may change over time and is not an appropriate dataset for shared research results. See available benchmark datasets if that is your intent.

It is very important for the analysis to get an overview of the dataset before and during the analysis in order to follow the right path towards a feasible analysis, good results and conclusions.

2.1 The Dataset

```
## Classes 'data.table' and 'data.frame':   9000061 obs. of  6 variables:
##  $ userId   : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ movieId  : num  122 185 231 292 316 329 355 356 362 364 ...
##  $ rating   : num  5 5 5 5 5 5 5 5 5 5 ...
##  $ timestamp: int  838985046 838983525 838983392 838983421 838983392 838983392 838984474
##  $ title    : chr  "Boomerang (1992)" "Net, The (1995)" "Dumb & Dumber (1994)" "Outbreak
##  $ genres   : chr  "Comedy|Romance" "Action|Crime|Thriller" "Comedy" "Action|Drama|Sci-Fi
##  - attr(*, ".internal.selfref")=<externalptr>
```

As you can see The dataset consist of 9,00,061 greetings and 6 variables. This means that it's a relatively large dataset and we will analyze the dataset firstly, secondly we will visualize the data on with the toolbox that we have obtained on these 8 courses.

```
##      userId         movieId          rating        timestamp
##  Min.   :    1   Min.   :    1   Min.   :0.500   Min.   :7.897e+08
##  1st Qu.:18122   1st Qu.:  648   1st Qu.:3.000   1st Qu.:9.468e+08
##  Median :35743   Median : 1834   Median :4.000   Median :1.035e+09
##  Mean   :35869   Mean   : 4120   Mean   :3.512   Mean   :1.033e+09
##  3rd Qu.:53602   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:9000061     Length:9000061
##  Class :character   Class :character
##  Mode  :character   Mode  :character
##
##
##
```

2.2 Summary of the validation set

```
##      userId         movieId          rating        timestamp
##  Min.   :    1   Min.   :    1   Min.   :0.500   Min.   :7.897e+08
##  1st Qu.:18127   1st Qu.:  653   1st Qu.:3.000   1st Qu.:9.468e+08
##  Median :35719   Median : 1835   Median :4.000   Median :1.036e+09
##  Mean   :35878   Mean   : 4121   Mean   :3.512   Mean   :1.033e+09
##  3rd Qu.:53649   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:999993      Length:999993
##  Class :character   Class :character
##  Mode  :character   Mode  :character
##
##
##
```

Chapter 3

Data cleaning and wrangling

Before the analysis we need to clean the data and this encompasses:

- Wrangling
- joining
- Eliminating columns
- Save the new file with the new corrections

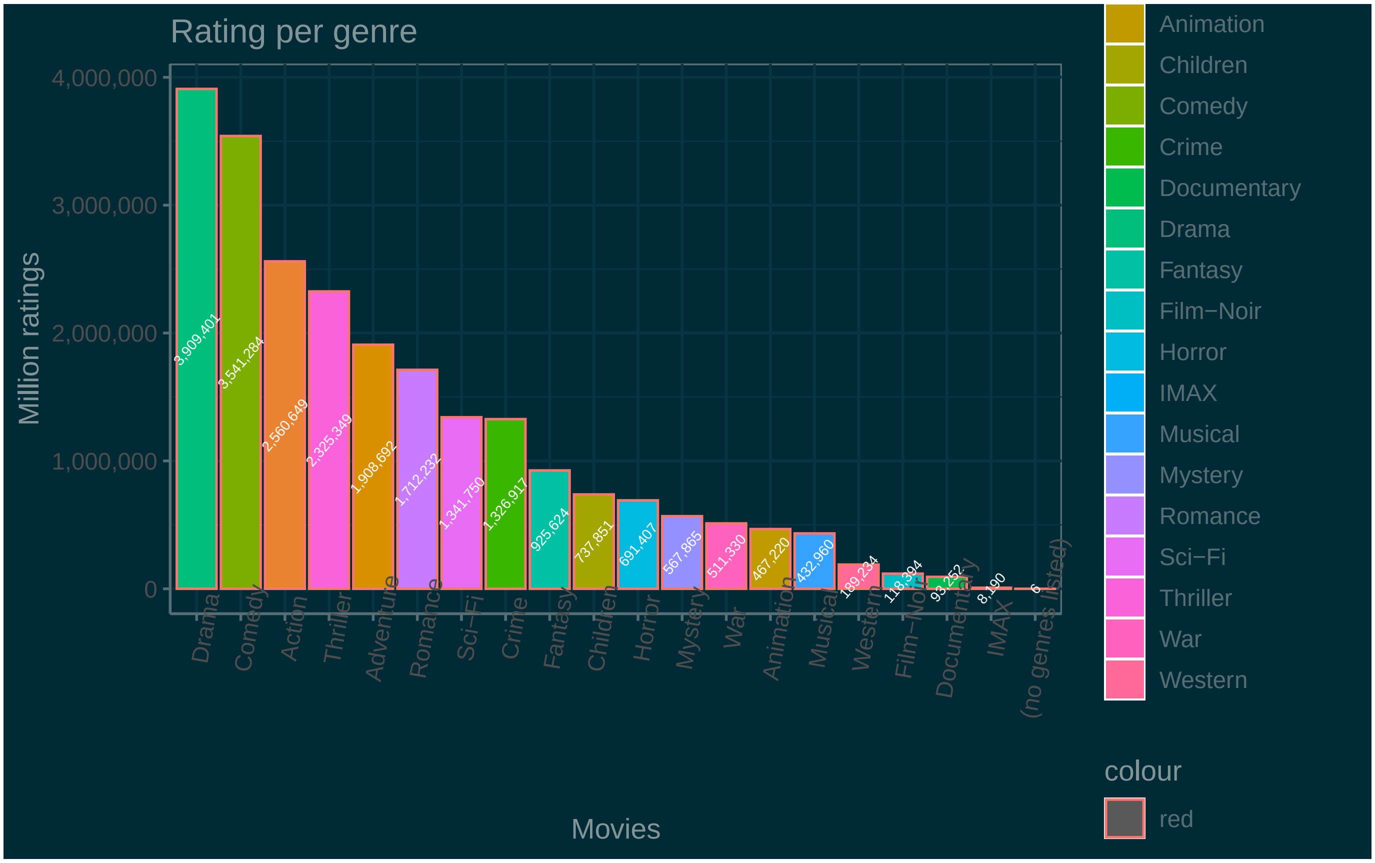
Most of the data has be cleaned, and in practice this is often the hardest and the most boring part.

Chapter 4

Visualisation of the training set

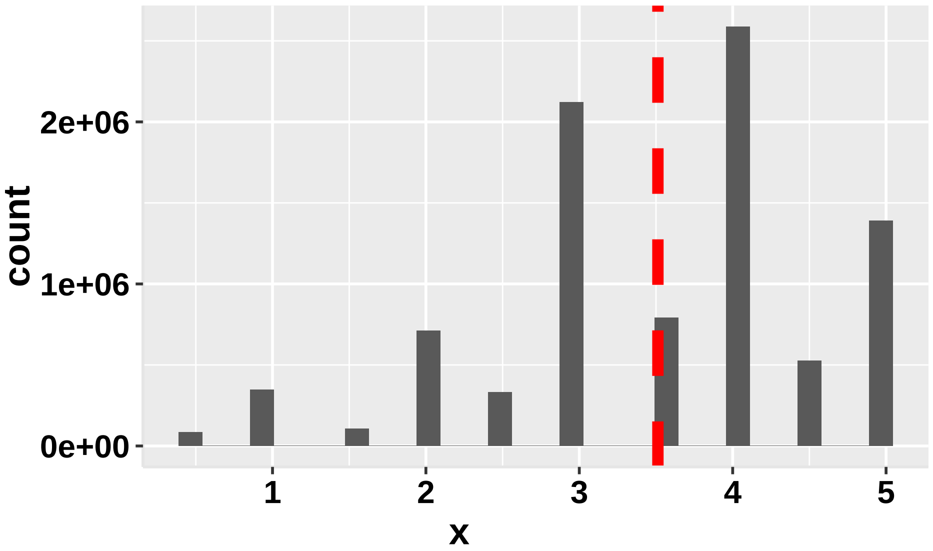
To show the force and flexibility of R studio, the visualization tools are excellent. All though a bit complicated to learn. the improve the overview of the dataset and the rating, the table below show the rating distributed on genres.

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



If you sort the genres in total rating descending order you con conclude that **drama** top with 3.9 million rating point with comedy as the 3.5 million runner up.

To improve the knowledge of interrelations of the viewers preferences the segment that prefer drama might also prefer action and the segment that loves comedy also prefer romance. That would be an obvious choice of segmentation of the population.



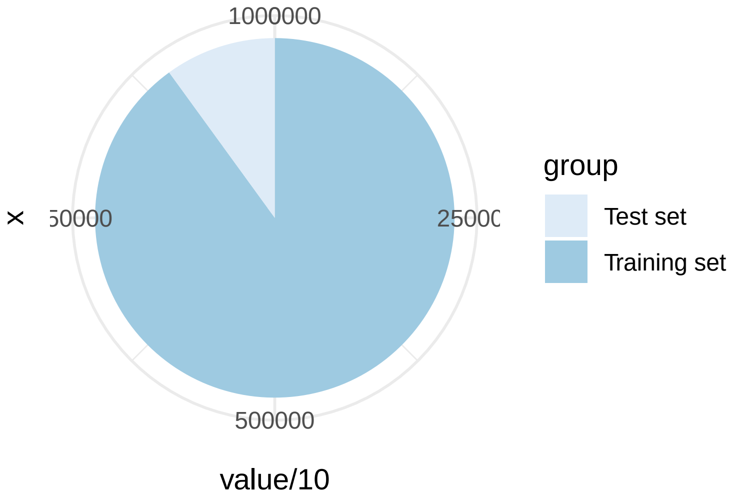
In the histogram above you may notice that the distribution on ratings is visualized. the highest average score is 4 and the mean is the dashed line on the value 3.51 ratings.

Chapter 5

Training set and test set

We have now divided the dataset into the training set and the tests. These two data sets will be used for analyses and developing the machine learning model to predict the future ratings.

The training set consists of 9.000,061 and the test set of 999,993 observations. 11.11% sample of the population. This is an large sample and I expect to predict future ratings with a good accuracy.



Chapter 6

Building a model of Machine Learning

The model which are needed for the training the dataset will be the following:

$$\sqrt{b^2 - 4ac}$$

6.1 Method 1

The simplest model is to use the average across every user and every movie as all of our predicted ratings. This model follows the citation,

$$Y_{u,i} = \mu$$

where $Y_{u,i}$ is the predicted rating of user u and movie i and μ is the average rating across all entries. This is computed as 3.512 (`mean(edx$rating)`). This is shown in a histogram.

```
## [1] 1.060651
```

6.2 Method 2

In order to improve the model, I add an independent error term $b_{u,i}$ that expresses rating differences for users and movies.

The improved model is:

$$Y_{u,i} = \mu + b_i$$

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

```
## [1] 0.9437046
```

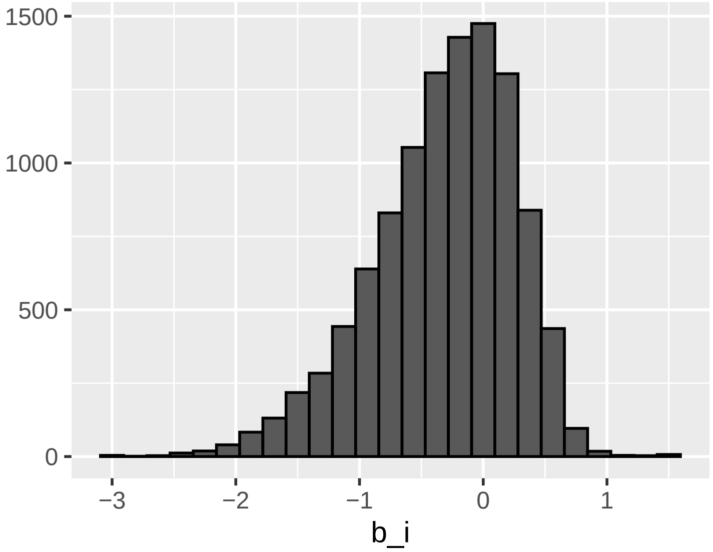


Figure 6.1: Here is shown the frequency of xxx

6.3 Method 3

The user bias term b_u will be used to reduce the errors additionally. The addition shall reduce the variability. Each user u is given a bias term take into account their predicted movies.

Then I have:

$$Y_{u,i} = \mu + b_i + b_u$$

```
## [1] 0.8655329
```

6.4 Method 4

The last thing I will implement is minimizing the big errors in my predictions. Regularization recudes the bias of sample size too small

For instance, our b_i term accounts for the average deviation on all ratings of a movie, whether there are 5 or 50 ratings on the film.

This can be seen in the following citation:

$$N \sum_{u,i} (Y_{u,i} - \mu - b_i - b_u)^2 + \lambda(\sum_i b_i^2 + \sum_u b_u^2)$$

Minimizing the biases using a single λ is the goal to our model shown above. The following test I use `lamda <- seq(from=0, to=10, by=0.25)`.

```
## [1] 0.8649857
```

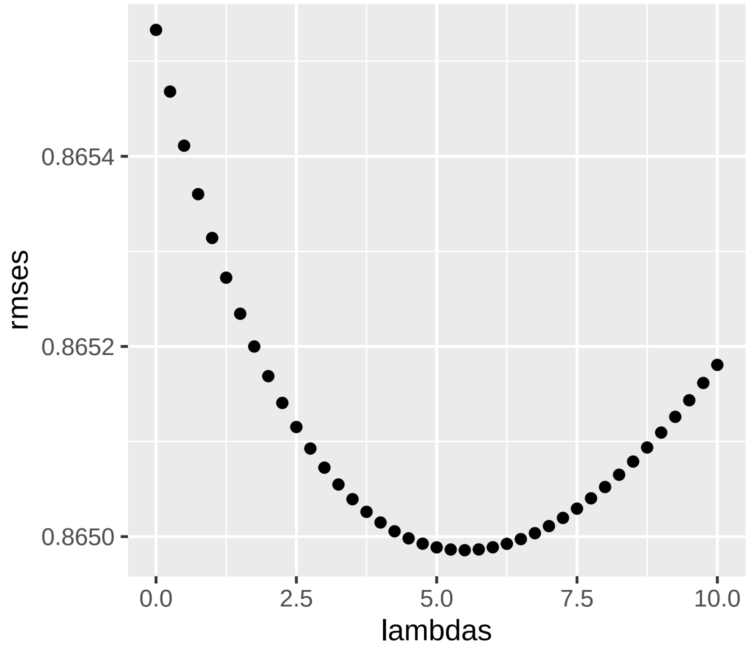


Figure 6.2: Shows the RMSEs of each *lambda* - seq(from=0, to=10, by=0.25)

6.5 Final

Now we introduce the user bias term b_u in order to further improve our model. This term minimizes the effect of extreme ratings made by users that love or hate every movie. Each user u is given a bias term that sways their predicted movies. Our updated model is:

$$Y_{u,i} = \mu + b_i + b_u$$

```
## [1] 0.8649857
```

Chapter 7

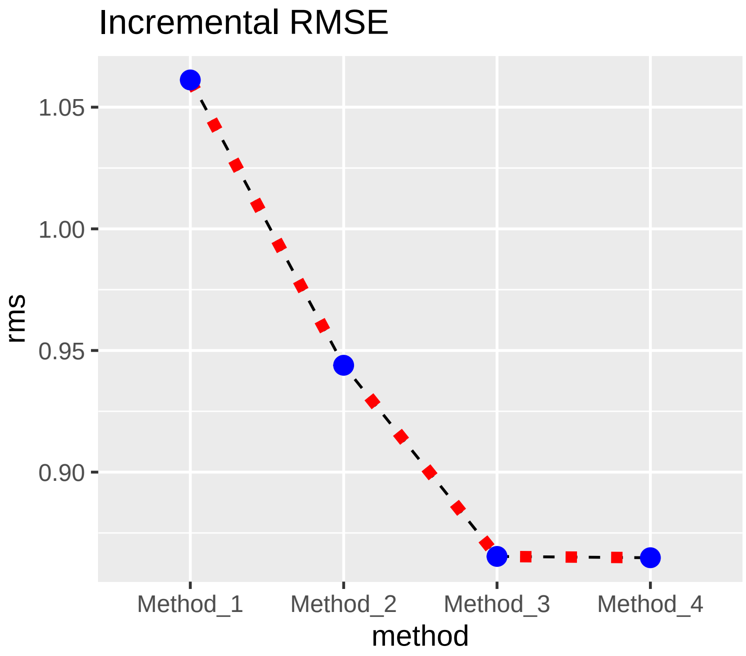
Results and conclusion

With the term web scraping the model could be improved by adding additional dimensions into the datasets such as expected costs of movie, rescaling ratings etc.

During the process of optimizing the models it seems obvious that movie and userid explain more than the genre. After training different models, it's very clear that movieId and userId contribute more than the genre predictor.

To get an overview of the obtained and calculated RMSEs please note the following table.

Method	RMSE
Method 1	1.06120
Method 2	0.94391
Method 3	0.86535
Method 4	0.86481



My take-out from the course The knowledge that i have gained before this course - was started in 1990 with statistics with pen and paper, so I have refreshed and updated my statistics, but most importantly - R - statistics calculations, R Markdown and the power of the new technology, i can certainly relate to now.

It has been a very interesting journey from my perspective. The core for me is that it is very important to improve the communication of knowledge, the visual part. That's a science. Its far more important than the quantity of charts that you may have in one report.

The idea must be to make complex relations into simple keystrokes for the audience or target group.

January 2021

Jan Thomsen

Chapter 8

Appendix

8.1 1 - Acknowledgements

My wife for accepting my attendance on these 9 courses