MovieLens Project

Aleksandra Sledziewska

6/8/2019

Introduction

The aim of this project is to create a movie recommendation system based on the MovieLens dataset. The system should find movies appropriate for a given user that he or she has not seen yet. The system will predict a rating that a user would give to a chosen film.

The data used in the project is just a subset of a dataset with millions of ratings. The dataset contains information about movie ID, user ID, movie title, given rating, timestamp and genres of the movie.

```
##    userId movieId rating timestamp
## 1       1     122      5 838985046
## 2       1     185      5 838983525
## 4       1     292      5 838983421
## 5       1     316      5 838983392
## 6       1     329      5 838983392
## 7       1     355      5 838984474
## 8       1     356      5 838983653
## 9       1     362      5 838984885
## 10      1     364      5 838983707
## 11      1     370      5 838984596
##                                    title
## 1                          Boomerang (1992)
## 2                            Net, The (1995)
## 4                            Outbreak (1995)
## 5                            Stargate (1994)
## 6              Star Trek: Generations (1994)
## 7                    Flintstones, The (1994)
## 8                        Forrest Gump (1994)
## 9                    Jungle Book, The (1994)
## 10                     Lion King, The (1994)
## 11 Naked Gun 33 1/3: The Final Insult (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
## 8                    Comedy|Drama|Romance|War
## 9                  Adventure|Children|Romance
## 10 Adventure|Animation|Children|Drama|Musical
## 11                           Action|Comedy
```

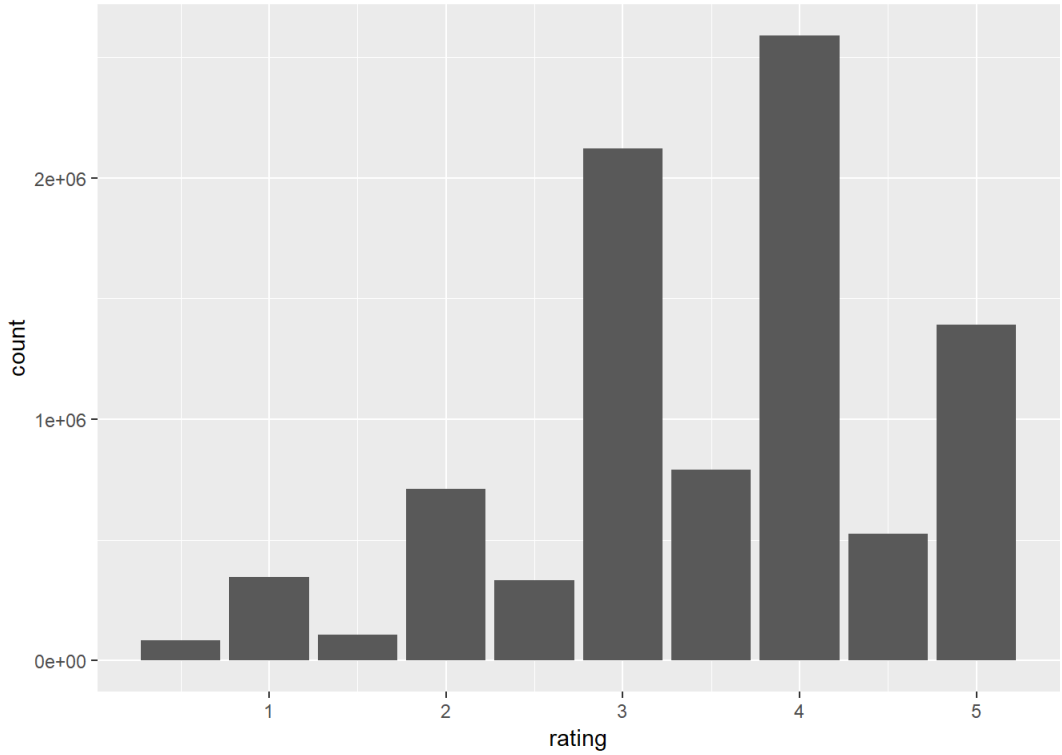
First of all, the dataset was examined and it was split into training and test set. Then, basic visualizations were created to better understand the data. Finally, the recommendation system was built step by step, starting from simple average rating and adding additional effects. The system was assessed using RMSE calculated on the test set.

Methods

Firstly, the dataset was split into training and test set. The test set accounted for 10% of the original dataset. The training set contained over 9 M observations of 6 variables.

Secondly, through data visualization the data was checked for outliers and mistakes. The distribution of given ratings is presented on a

graph below:



Given ratings ranged from 0.5 to 5 and users were more likely to give whole star ratings than half star ratings. Users most often gave 4 as a rating.

Most rated genres included Drama and Comedy, however, some movies were categorized into a few genres.

```
## # A tibble: 10 x 2
##    genres                    count
##    <chr>                     <int>
##  1 Drama                    733296
##  2 Comedy                   700889
##  3 Comedy|Romance           365468
##  4 Comedy|Drama             323637
##  5 Comedy|Drama|Romance     261425
##  6 Drama|Romance            259355
##  7 Action|Adventure|Sci-Fi  219938
##  8 Action|Adventure|Thriller 149091
##  9 Drama|Thriller           145373
## 10 Crime|Drama              137387
```

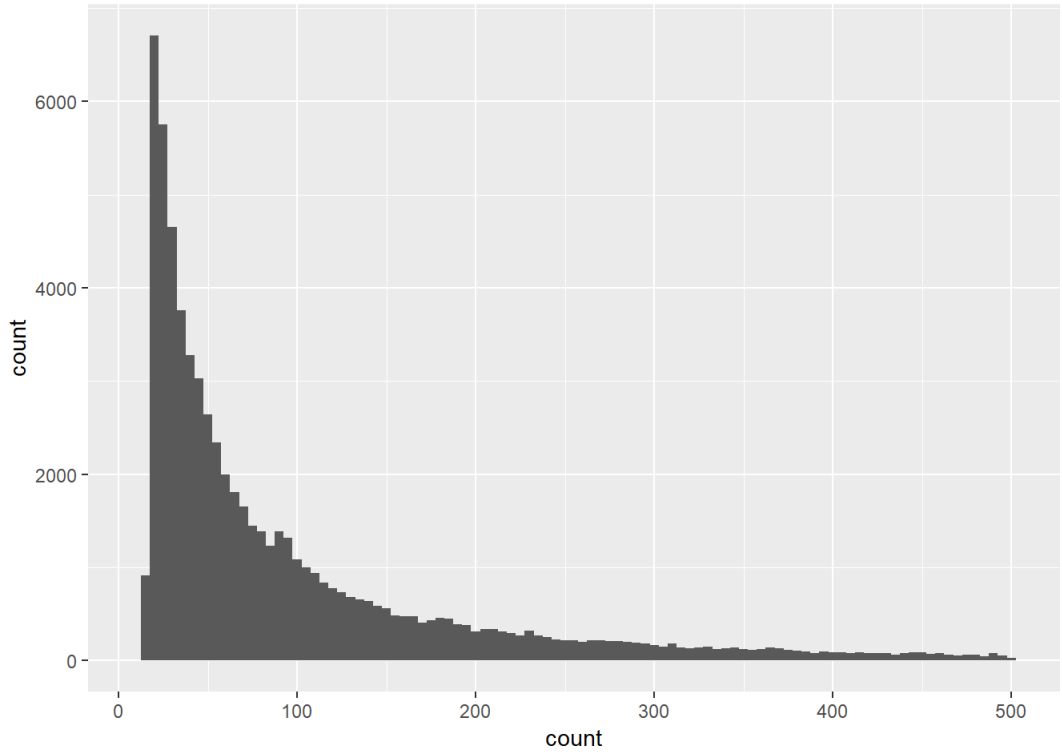
Users most often rated movies such as Pulp Fiction, Forrest Gump and The Silence of the Lambs.

```
## # A tibble: 5 x 3
## # Groups:   movieId [5]
##   movieId title                          count
##     <dbl> <chr>                          <int>
## 1     296 Pulp Fiction (1994)            31362
## 2     356 Forrest Gump (1994)            31079
## 3     593 Silence of the Lambs, The (1991) 30382
## 4     480 Jurassic Park (1993)           29360
## 5     318 Shawshank Redemption, The (1994) 28015
```

The most active users rated over 6000 movies.

```
## # A tibble: 5 x 2
##   userId count
##    <int> <int>
## 1  59269  6616
## 2  67385  6360
## 3  14463  4648
## 4  68259  4036
## 5  27468  4023
```

However, the majority of users rated fewer than 100 movies.



Finally, the predictive model was built in 4 steps:

- giving average rating (the same for all films)
- adding average rating for each movie (accounting for movie effect)
- adding average rating for each use (accounting for user effect)
- adding penalty for large estimates that come from small sample sizes (regularization)

This approach allows to see whether the model is improving with every consecutive step and by how much. In the first step every movie is given the average rating from the whole dataset. Then, new averages are calculated for each movie, as in general we have better movies and worse movies. Later, averages are calculated for each user as well, because some of the users are more critical, meanwhile others appreciate most of watched movies. Finally, a penalty is added not to overtrust some predictions coming from very small subsamples (e.g. movie which was rated only 3 times), as they are burdened with high errors.

Results

The table below presents the RMSEs for each model with the RMSE of the final model equal to 0.8648.

method	RMSE
Just the average	1.0612018
Movie Effect Model	0.9439087
Movie + User Effects Model	0.8653488
Regularized Movie + User Effects Model	0.8648177

As mentioned before, the final models includes a prediction based on general average rating, average rating for each movie, average rating for each user and penalty for overtrusting small subsamples.

The RMSE was calculated on a test set. The final model is burdened with 86.48% error in comparison with the true rating a user would give to a given movie. It is a huge improvement compared to the primary model that was burdened with more than 100% error.

Comparison of exemplary predicted ratings and true ratings is presented below.

```
##    rating     pred
## 1     5.0 4.264512
## 2     5.0 4.992708
## 3     5.0 4.385029
## 4     3.0 3.347321
## 5     2.0 4.232387
## 6     3.0 2.762940
## 7     3.5 3.970675
## 8     4.5 4.133987
## 9     5.0 4.276113
## 10    3.0 3.305501
## 11    3.0 3.646216
## 12    3.0 3.590914
## 13    3.0 3.676672
## 14    3.0 4.146608
## 15    3.0 3.483450
## 16    3.0 4.221137
## 17    3.0 3.750400
## 18    3.0 3.466594
## 19    4.0 4.284264
## 20    5.0 3.876010
```

Conclusion

This project intented to build a movie recommendation system based on the MovieLens dataset. The data was examined, cleaned and visualized. When assigning a predicted rating a user would give to a movie, the final model took into account an average rating in the dataset, an average rating given by a user, an average rating given to a movie and a penalty for large estimates coming from small sample sizes. Eventually, the error of the predicted rating compared to the true rating was 86.48%.

Moreover, more indepth knowledge about the user and his/her preferences would allow to build even better model. Some additional information about the user or the movie would be facilitative. For example, an information about leading actors in a movie could impact the user's rating (he/she can favor some actors and rate the movies higher). In addition, the average for each genre could be calculated too, probably some of the genres are rated higher than the others. However, some of the movies are classified to more than one genre, so it can be quite misleading.

The model attained in this project is satisfactory, nevertheless, there is still some potential for improvement and better understanding of a user's movie taste.