Project 1-Movielens

Haas Nicolas

30 11 2020

Introduction

Companies like Youtube or Netflix are quite famous because of their well performing movie/video recommendation systems.If a user only gives a small peace of information about his preferences (one movie that he likes), applied algorithms can well predict which movie he wants to look next. Nevertheless, the main challenge in building such algorithms/recommendations systems lies in finding an efficient way which is still consistent over time. In other words, it is important to find the main sources that account for the variation in the data so that at the end a still "general" model can be used to different people over different years. Hence, this analysis is interested in the question how fast can we get a useful model? Or in other words, how many predictors do we need to have to recommend/predict well?

This analysis therefore uses a data set which is a subset of the MovieLens data and has been published by the grouplens researcher group (https://grouplens.org/datasets/movielens/10m/). This data set contains about 10 million movie ratings given by different people all over the world and over different years but it is still highly unbalanced (Some users only rate one movie and vice versa). The dependent variable is the continuous rating variable (0.5,1,…,4.5,5.0). The main idea is to work with movie ID and user ID that can be used as predictors (independent variables) because we know that not all users have the same preferences and not all movies represent the same genre, have the same actors,etc. Thus, it is important to include user and movie specific preferences/effects. Though, not only movie and user specific effects will be used to look for variation in the data but also rating patterns which are more general and are not specifically related to users and/or movies. In this context, a difference between half star and full star ratings can also improve the algorithms at the end. The performance of the algorithms itself will be measured by the root mean squared error (RMSE). At the end, the algorithms will lead to a RMSE of about 0.8639 (data split - 10 % test and 90 % train set).

Raw Data

To understand the main challenge in building such a recommendation system it is a good idea to explain first why it is important to include movie and user specific effects. Below you can find some examples of the mean rating of some movies. There are huge differences in their mean rating. It lays therefore in the nature of the problem that not all movies have the same properties (different genres, different actors, different quality/budget, different landscapes,etc.). Hence, different movies will be rated differently simply because of the reason that they are different. Thus, it is important to account for differences in movies.

Table 1: Some Examples - Mean Of Rating By Movie X

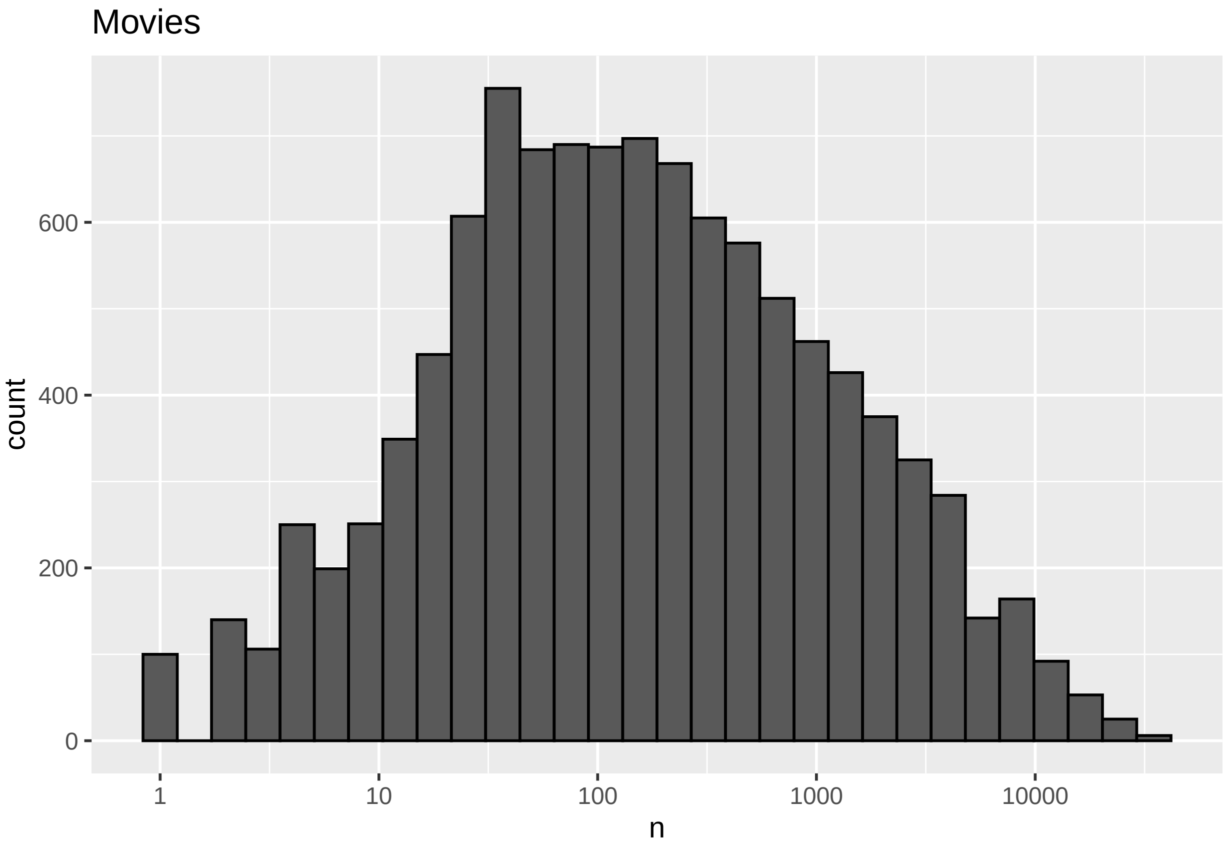
movieId	mean_rating
1	3.928769
5094	1.839175
6099	1.902778
7113	3.300000
8702	2.833333
32375	3.450000

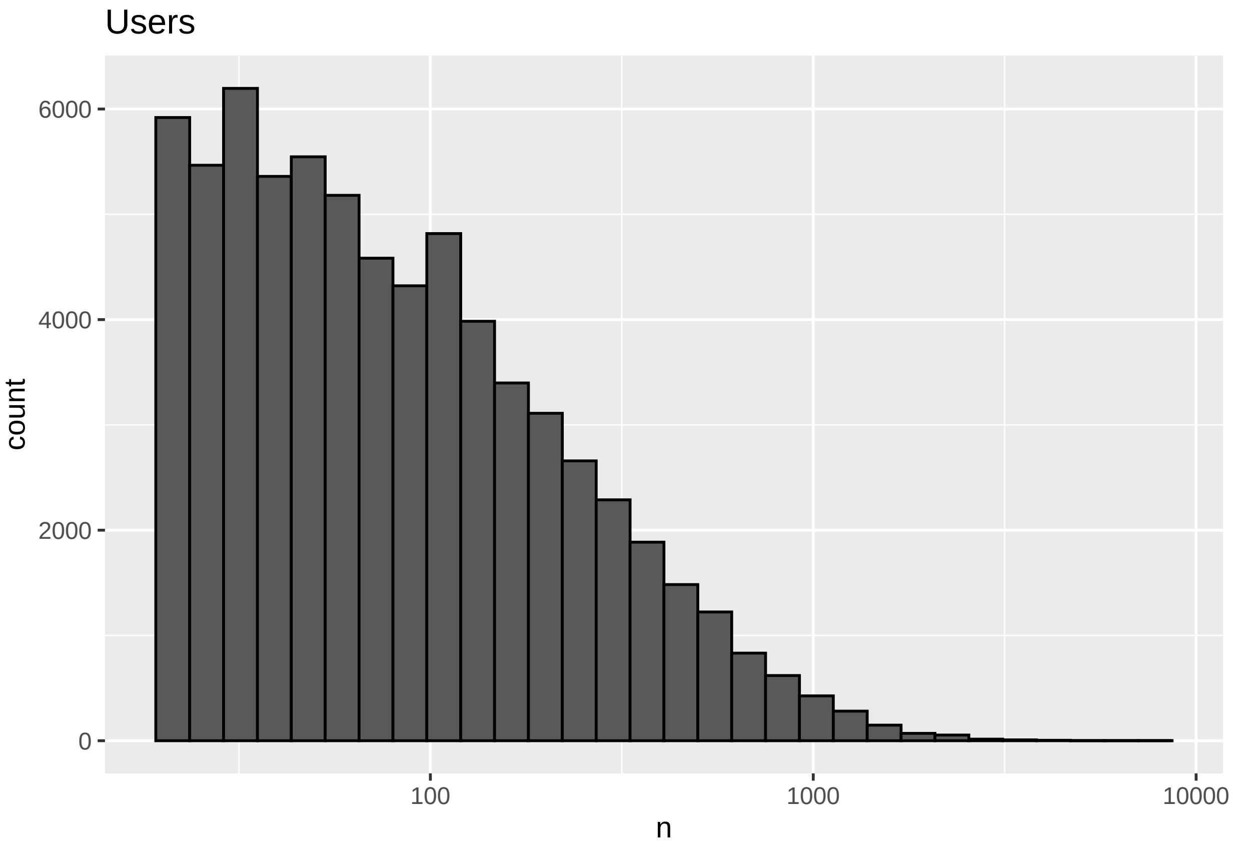
The same can also be observed for users when looking to their mean rating patterns. User differences can occur because of their different preferences, different nationality, different intelligence levels, etc. All this leads to the reason that not all users are the same. Some prefer comedies for example more than thrillers or some dislike french movies, etc. Therefore, it is also important to account for these differences in the mean rating pattern in terms of user specific effects.

Table 2: Some Examples - Mean Of Rating By User X

userId	mean_rating
1	5.000000
10279	3.581539
20537	3.322368
30770	3.771739
40986	3.704546
51168	3.921260
61349	3.519231

Furthermore, users do not only differ because of their preferences and movies do not only depend on their natural properties but there are also differences in the rating activities within movies and within users. Thus, Figure "Movies" and "Users" show the distribution of the rating activity by movies and users specifically. For example in "Movies" we can see that most movies are rated between 10 and 3000 times (Still a normal distribution). Only a small part of movies are rated less than 5 times or by more than 10.000 (more outliers). A bit more differently looks the distribution of "Users". The part of users with only a few of ratings seems to be evenly distributed to the part of users which already have some experience (close to 100 rated movies). Only after reaching this critical point of 100 rated movies, users begin rapidly to decrease. In general, these two figures already show that there is specific rating activity for every user and for every movies. Furthermore, the number of users with very little experience (very few observations) is much larger compared to the movie effect. In other words, users with very few ratings gain much more weight in the data set compared to movies with very few ratings. The question then is, shall we trust these inexperienced users? Probably not and this analysis will account for this problem in terms of penalty term that we will see later.


Movies
count
600
400
200
0
1
10
100
1000
10000
n


Users
6000
4000
2000
0
count
100
1000
10000
n

As already mentioned in the introduction, it is also useful to look to some general effects that might “bias” the results. The word “bias” is often a sample property or a sample problem (Sample does not entirely represent the true population). In the context of this analysis, we look to the difference between half and full star ratings. Table “General Bias - Half(0) Compared To Full(1) Star Rating” shows that there is a clear discrepancy in the number of observations and their mean value between half and full star ratings in the data. Why can this be? First, we can once again get the impression that it depends on users because they probably don‘t know that they can also give a half star rating. But we will see later, that after including user specific effects, the difference still remains so that it is more a general problem. A second reason therefore could be that it is for example related to general rating terms/conditions and sometimes they apply and sometimes they do not apply. We can here think about a yearly cut in the rating terms/conditions. For example, in year XXXX a half star rating has been introduced and therefore it was not possible to give a half star rating before. This could explain the huge gap in observations between half and whole star ratings. Despite the fact that it is not clear why there is a discrepancy in observations and the mean value between half and full star ratings, it seems to be a sample issue (for example related to year of the rating). Therefore it is important to take this variation in this analysis specifically into account.

Table 3: General Bias - Half(0) Compared To Full(1) Star Rating

wholestar	number_of_observations	mean_value
0	2048230	3.350273
1	7951824	3.554188

Model/Method

With all this insights in mind gained by the data exploration, the analysis looks for the following model (See also Irizarry, 2019):

rating(i,u) = a + b(i) + b(u) + c + Error_term(i,u)

where a:= the overall mean of the general rating pattern

and b(i):= Different coefficient/Mean value for every movie

and b(u):= Different coefficient/Mean value for every user

and c:= Different coefficient/Mean value for half and whole star ratings

and Error_term(i,u):=independent error term for every movie and every user

The idea is to measure the user (movie) specific effects by the mean of all ratings given by each user (for each movie). Later, the model is adapted by a penalty term for specifically user effects because of the mentioned large amount of inexperienced raters that has already been mentioned in the data visualization. The penalty term takes place in the averaging effect of every users overall ratings in terms of parameter lambda. Because of this effect, a single rating will not only be out weighted by the total number of ratings of this user but also by a parameter lambda. Hence, this effect becomes precisely then large when a user only rates a couple of movies (or only just one). In general we can use the following adaptation:

b(u)_hat = 1/(n(u) + lambda) * sum(rating - (a + b(i))

where n(u):= user specific sample size

and lambda:= penalty parameter

b(u)_hat;= estimated values of user specific effects

All other components have already been explained above. The choice of the value of lambda will be calculated by cross validation and leads to an optimal choice of lambda equal to 5.5.

NB: We will later see that the analysis also includes a penalty term for movies but the decrease in RMSE is only small (b(i)_hat = 1/(n(u) + lambda)*sum(rating - a).

Results

Despite the rough elaboration of the final model in the last chapter, the analysis itself takes a more step-wise process. You can see below the results of a model that neither includes a penalty term nor a a variable for whole star ratings. The first model includes only the overall mean as regressor, the second one adds a movie specific effect and the third model also adds the user specific effect. Because the accuracy (measured by the RMSE) is increasing by each adaptation (RMSE goes down), the steps chosen are reasonable.

method	RMSE
Model 1 - Only Overall Variation	1.0612018
Model 2 - Movie Effect Model	0.9439087
Model 3 - Movie And User Effects Model	0.8653488

Nevertheless, to see if the model is also doing well in detail we can look to the largest errors/residuals when movie and user specific effects are included. Table "Ten Largest Mistakes When Using User And Movie Specific Effects Without Penalty Term" clearly shows that the largest mistakes are reasonable. For example a block buster movie like Lord of the Rings I (awarded by several golden globe prices) has a mistake given by -4.70 which means that it has a high predicted rating (high mean rating) and gets only such a large error because there were some few users (probably one) who dislike this movie. All other movies of this table can be explained in the same way. Therefore our model already performs well without penalty term incorporation or without the use of a whole star variable.

Table 5: Ten Largest Mistakes When Using User And Movie Specific Effects Without Penalty Term

title	residual
Godfather, The (1972)	-4.933196
Casablanca (1942)	-4.793064
Lord of the Rings: The Two Towers, The (2002)	-4.707729
Christmas Story, A (1983)	-4.671677
Christmas Story, A (1983)	-4.583300
Searching for Bobby Fischer (1993)	-4.491043
Never on Sunday (Pote tin Kyriaki) (1960)	-4.453936
Night on Earth (1991)	-4.432325
Duck Soup (1933)	-4.418881
O Brother, Where Art Thou? (2000)	-4.399042

Nevertheless, let‘s look whether the model can perform better in including the mentioned penalty term for user and movie effects and also the whole star variable. Looking to the table below, we can always see an improvement although the impact of including a penalty for movies is the smallest one (Improvement only by 0.00016 units from model 4 to model 5). This stays in accordance with the insights gained by the raw data elaboration. Whereas movies only has a small number of outliers (few movies were rated seldom or very often), the impact of inexperienced raters on the sight of users is much larger. This explains the only small win in accuracy from model 4 (RMSE: 0.86497) to model 5 (RMSE: 0.86481). Furthermore, the incorporation of the whole star variable also has its desired effect, the RMSE goes down to 0.8639.

method	RMSE
Model 1 - Only Overall Variation	1.0612018
Model 2 - Movie Effect Model	0.9439087
Model 3 - Movie And User Effects Model	0.8653488
Model 4 - Movie And User Effects Model with Regularization Only For User	0.8649710
Model 5 - Movie And User Effects Model with Regularization For User And Movie	0.8648178
Model 6 - Movie And User Effects Model With Regularization For User And Movie, Additionally Wholestar Variable	0.8639453

A last verification of the residuals shows that the predicted mistakes are still reasonable (Almost the same list as seen before but the values are slightly changing-the errors get smaller). The final model (Model 6) performs best and is therefore the most preferred model of this analysis.

Table 7: Ten Largest Mistakes When Using User and Movie Specific Effects + Penalty Term + Whole Star Variable

title	residual
Godfather, The (1972)	-4.817171
Christmas Story, A (1983)	-4.586440
Lord of the Rings: The Two Towers, The (2002)	-4.581509
Casablanca (1942)	-4.559835
Christmas Story, A (1983)	-4.441251
Never on Sunday (Pote tin Kyriaki) (1960)	-4.347812
Duck Soup (1933)	-4.333499
Beauty and the Beast (1991)	-4.315912
Manchurian Candidate, The (1962)	-4.281930
Searching for Bobby Fischer (1993)	-4.280254

Conclusion

Building a recommendation system for movies is far from being easy because we have different users (different preferences) and different movies (different genres, different actors, etc.). Furthermore, not all movies and users represent the same rating activity. Although the data are very handy in this data set, we could directly observe that there are different patterns in the rating activity between movies and users. Whereas movies tend to be more normally distributed around a certain amount of ratings, the same is not true for users. The shape of the user curve was much more left skewed and therefore also users with few ratings (inexperienced users) gain much more weight. The solution to this was the incorporation of a penalty term that has the desired effect only when the number of ratings per user are very small. The result directly improved in accuracy (measured by the RMSE) even when this penalty term was used for both, movies and users. The last step then added a whole star variable which differed between half and whole star raters. The large amount of whole star rater (observed in the data visualization) already let assume that there is an abnormality in the data and we should control for this effect in this analysis (as a control variable). Fortunately, it decreased the RMSE still by some units. Despite the fact, that the reason for this difference between half and whole star raters are not clear for the moment (Different rating conditions over the years?), the incorporation of this variable certainly decreased the variation in the model. The final result of the RMSE was close to 0.8639.

When we talk about limitations, we can still keep going with the last mentioned point. The weakness of the preferred result is that we still do not know why there is a difference between half and whole star rater? Because this was a general bias (user and movie specific effects already incorporated), it seems to depend more on years or another factor. Probably there was a cut in the rating terms in one year. This could explain that rating patterns in general are a bit higher or lower compared to another year, simply because of rating condition changes over the years. A second weakness, is that this is still a robust model but not really precise. We only included specific effects for users and movies and the whole star variable but there are still differences by genre, by an interaction between movies and users, by years of ratings, etc. Therefore the model can be applied in general but has to be adapted in cases where the population is quite specific. For example, when you want to apply this model to a country specific recommendation system (for example for television), it has to be adapted once again (Probably because of national movie and actor bias). A third weakness is the data set itself. For example we don‘t have many user information. At least the address of the user could give some important information or the year when the rating was given. Therefore, future research can not only be done in the model improvement (more factor analysis) but also in data collection. More user information is an important point in this aspect.

Finally, to answer the question of the introduction how fast we can get a relatively precise model, this analysis could give robust results in only accounting for an overall, a user specific and a movie specific effect, adapted for a penalty term for outliers in the user and movie rating activity. In the end, the whole star variable also helped to reduce the RMSE quite efficiently (few computation time with large precision effect) but one can assume that this was only a sample problem specifically related to the data set. Hence, a well performing recommendation system certainly depends on the sample variation (national movie recommendation system for television would certainly look differently). Though, the variables chosen to perform the final model in this analysis already help to explain a large proportion of the variation in a large movie recommendation model (with a large set of movies).

References

Rafael A. Irizarry (2019), Introduction to Data Science: Data Analysis and Prediction Algorithms with R

https://www.edx.org/professional-certificate/harvardx-data-science

https://movielens.org/

https://grouplens.org/datasets/movielens/10m/