Movie Lens

Carlos Felipe Rengifo Rodas

2022-04-04

Contents

1 Introduction

The MovieLens dataset used in this project comprises 10 millions ratings for 10677 movies rated by 69878 users. The scale for the rating goes from 0 to 5 with increments of 0.5. The product between the number of users and the number of movies is close to 746 millions, which is greater than the number of ratings. This difference indicates that not every user rated every movie. The MovieLens dataset has six columns which are: (1) the movie identifier, (2) the user identifier, (3) the rating of the movie, (4) a time stamp with the date of the rating, (5) the title of the movie, and (6) the genres of the movie.

The purpose of this project is to build a linear model to predict the rating $r_{m,u}$ that the user u will give to the movie m. Such a model can be mathematically written as follows:

$$r_{m,u} = \mu + \lambda \; (b_m + b_u) + \epsilon_{m,u}$$

μ is the average movie rating, b_m represents the movie-to-movie variation, b_u is the user-to-user variation, $\epsilon_{m,u}$ is a zero mean random variable representing uncertainty, and λ is a penalty factor that avoids large values for b_m and b_u. A second model that we want to explore includes the movie genre effect b_g:

$$r_{m,u,g} = \mu + \lambda \; (b_m + b_u + b_g) + \epsilon_{m,u,g}$$

Table 1: Most rated movies

Title	Num. of ratings
Pulp Fiction (1994)	34864
Forrest Gump (1994)	34457
Silence of the Lambs, The (1991)	33668
Jurassic Park (1993)	32631
Shawshank Redemption, The (1994)	31126
Braveheart (1995)	29154

2 Methods

2.1 Data exploration and visualization

The Table 1 presents the titles of the most rated movies. The Figure 1 shows a histogram with the distribution of ratings among movies, and the Figure 2 shows a histogram with the distribution of ratings among users.

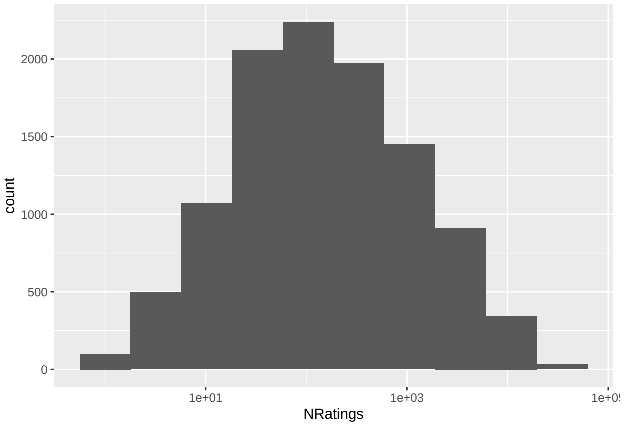

Figure 1: Distribution of ratings among movies

2.2 Insights gained

The data exploration and visualization of the previous section permit us to conclude that not every movie received the same number of ratings, and that no every user rated the same number of movies.

2.3 Modeling approach

The approach followed in this report consists of building models of increasing complexity until a mean square error of less than 0.86490 is obtained. These are the models that will be used:

1. $r_{m,u} = \mu + \epsilon_{m,u}$.
2. $r_{m,u} = \mu + b_m + \epsilon_{m,u}$.
3. $r_{m,u} = \mu + \lambda\, b_m + \epsilon_{m,u}$.
4. $r_{m,u} = \mu + \lambda\, (b_m + b_u) + \epsilon_{m,u}$
5. $r_{m,u,g} = \mu + \lambda\, (b_m + b_u + b_g) + \epsilon_{m,u,g}$

The quality of a model will be assessed according to the root mean squared error, which is defined as follows:

$$RMSE = \sqrt{\sum_{k=1}^{N} (r_{m,u} - \hat{r}_{m,u})^2}$$

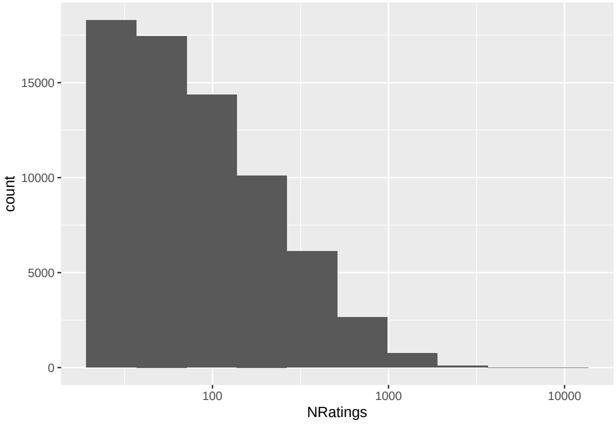

Figure 2: Distribution of ratings among users

$r_{m,u}$ is the actual rating given to movie m by the user u, and $\hat{r}_{m,u}$ is the prediction of such rating according to one of the models previously described. The lower the RMSE, the higher the accuracy of a model.

2.4 Data cleansing

2.4.1 Download the MovieLens database

The first step to build the dataset is to download the file *ml-10m.zip* from the MovieLens website. *ml-10m.zip* contains the files *ratings.dat* and *movies.dat*. The rows of the first file comprise a user identifier, a movie identifier, a rating, and a time stamp; while the rows of the file are composed by a movie identifier, a movie title, and a movie genre. The information obtained from *ratings.dat* and *movies.dat* is combined into a single data frame named *movieLens*.

2.4.2 Data partition

The second step is to use the function *createDataPartition* of the *caret* package to divide the data set into 90% for modelling and 10% for validation. The modelling and validation data frames will be termed *edx* and *validation*, respectively.

3 Results

3.1 Estimation of the population mean

The population mean is estimated by averaging all ratings of the *edx* data frame. The resulting value is $\hat{u} = 3.51$, which means that no matter the user or the movie, the predicted rating is:

$$\hat{r}_{m,u} = 3.51$$

The RMSE for this model is 1.0612, which means that in average our predictions differ in more than one star from the actual rating.

3.2 Estimation of the movie effect

The following model accounts for the movie effect:

$$r_{m,u} = \mu + b_m + \epsilon_{m,u}$$

Table 2: Movies with the highest average ratings

Title	Av. Rating	Num. Ratings
Hellhounds on My Trail (1999)	5	1
Satan's Tango (SÃ¡tÃ¡ntangÃ³) (1994)	5	2
Satan's Tango (SÃ¡tÃ¡ntangÃ³) (1994)	5	2
Shadows of Forgotten Ancestors (1964)	5	1
Fighting Elegy (Kenka erejii) (1966)	5	1
Sun Alley (Sonnenallee) (1999)	5	1
Blue Light, The (Das Blaue Licht) (1932)	5	1

b_m, which represent the movie bias, is estimated by averaging the ratings for the movie m, and then subtracting the overall mean μ. Since there are 10677 movies, the above model requires to estimate 10677 bias (i.e. $m = 1, 2, \ldots 10677$)

$$\hat{b}_m = \frac{1}{N_m} \sum_{u \in u(m)} r_{u,m} - \mu$$

N_m is the number of users who rated the movie m, and $u(m)$ is the set of users who rated such movie. The RMSE for this model is 0.9439. However, some values of $\hat{b}_m$ are not reliable because they were estimated using just a few ratings (one or two in some cases). The Table 2 shows that the best rated movies have only one or two ratings.

3.3 Estimation of the movie effect with regularization

Regularization permit us to take into consideration that the values of $\hat{b}_m$ estimated with very few ratings are random variables with higher variability than the $\hat{b}_m$ estimated with hundreds or thousands of ratings. In a model with regularization, a penalty factor λ multiplies the estimate that is regularized:

$$r_{u,m} = \mu + \lambda\, b_m + \epsilon_{u,m}$$

The regularized estimate, b_m in the present example, is estimated as follows:

$$\hat{b}_m = \frac{1}{N_u + \lambda} \sum_{u \in u(m)} r_{u,m} - \mu$$

When λ is varied from 0 to 10 with increments of 0.25, the value that minimizes the RMSE on the training set is $\lambda = 2.5$, and the corresponding RMSE is 0.9439. This RMSE is lower than the obtained from the previous model.

3.4 Estimation of the movie and user effects with regularization

The following model accounts for both the movie and the user effect:

$$r_{u,m} = \mu + \lambda(b_m + b_u) + \epsilon_{u,m}$$

b_u, which represent the user bias, is estimated by averaging the ratings given by user u, and then subtracting $\mu + b_m$. Since there are 69878 users, the above model requires to estimate 69878 bias (i.e. $u = 1, 2, \ldots 69878$)

$$\hat{b}_u = \frac{1}{N_u + \lambda} \sum_{m \in m(u)} r_{m,u} - \mu - b_m$$

Table 3: RMSE by model

Method	RMSE
u+e	1.0612018
u+bm+e	0.9439087
u+lambda*bm+e	0.9438521
u+lambda*(bm+bu)+e	0.8648170
u+lambda*(bm+bu+bg)+e	0.8644501

N_u is the number of movies rated by the user u, and $m(u)$ is the set of movies rated by the user u. When λ is varied from 0 to 10 with increments of 0.25, the value that minimizes the RMSE on the training set is $\lambda = 5.25$, and the corresponding RMSE is 0.8648.

3.5 Estimation of the movie, user and genre effects with regularization

The following includes the movie, user and genre effects:

$$r_{u,m,g} = \mu + \lambda(b_m + b_u + b_g) + \epsilon_{u,m,g}$$

b_g, which represent the genere bias, is estimated by averaging the ratings given to the movies of genre g, and then subtracting $\mu + b_m + b_u$. Since there are 797 genres, the above model requires to estimate 797 bias (i.e. $g = 1, 2, \ldots 797$)

$$\hat{b}_g = \frac{1}{N_g + \lambda} \sum_{m \in m(g)} \left[\sum_{u \in u(g)} r_{m,u,g} - \mu - b_m - b_u \right]$$

N_g is the number of movies of genre g, $m(g)$ is the set of movies of genre g, $u(g)$ is set of users who rated movies of genre g. When λ is varied from 0 to 10 with increments of 0.25, the value of λ that minimizes the RMSE on the training set is $\lambda = 5$, and the corresponding RMSE is 0.86.

3.6 Summary of results

The Table 3 summarizes the RMSE obtained from each model.

4 Conclusion

This report presented five models to predict the rating that a user gives to a movie. The complexity of the models is increasing. The first model considered only the overall mean rating, the second model comprised the overall mean rating and the movie effect, and the third model included the overall mean rating and the movie effect with regularization, which is necessary to avoid large values of the movie bias generated by movies with very few ratings. The fourth model considered the overall mean and regularized terms for the movie and the user effect. The five model was an extension of the fourth model that considered the effect of the movie genre in the rating. However, to include the genre effect decreased the RMSE from 0.8648 to just 0.8645. The main limitation of this report is that important predictors such as the product between user and genre of the movie were not included in the model. Another limitation of our study is that the penalty factors for the effect of user, movie and genre were the same. Future works should consider to include products between variables and apply multivariable optimization to obtain three regularization factors instead of one. We also hypothesized that including users' gender, geographic location, native language, and date of rating in the prediction model could improve accuracy.