MovieLens Project

Livio Catenazzo

6/01/2019

1.Introduction

Background and goals

The goal of this project is to create a movie recommendation system. This is a model with the power to predict movie ratings based on movies and users. Recommendation systems are a subclass of information filtering systems that seeks to predict the preference a user would give to an item. They are primarily used in commercial applications such Netflix, YouTube and Spotify.

The basic idea of a recommendation system is to give a helpful recommendation based on available data. To be more specific, the task is to predict the rating a particular user would give to a specific movie and therefore to provide matching movie suggestions to that user.

The RMSE will be used to evaluate how close predictions are to the true values.

Data available

The available data is the 10M version of the MovieLens dataset. This is dataset was released in Released 1/2009 and contains 10 million ratings and 100,000 tag applications applied to 10,000 movies by 72,000 users. It can be downloaded at the following link: http://files.grouplens.org/datasets/movielens/ml-10m.zip

2.Analysis

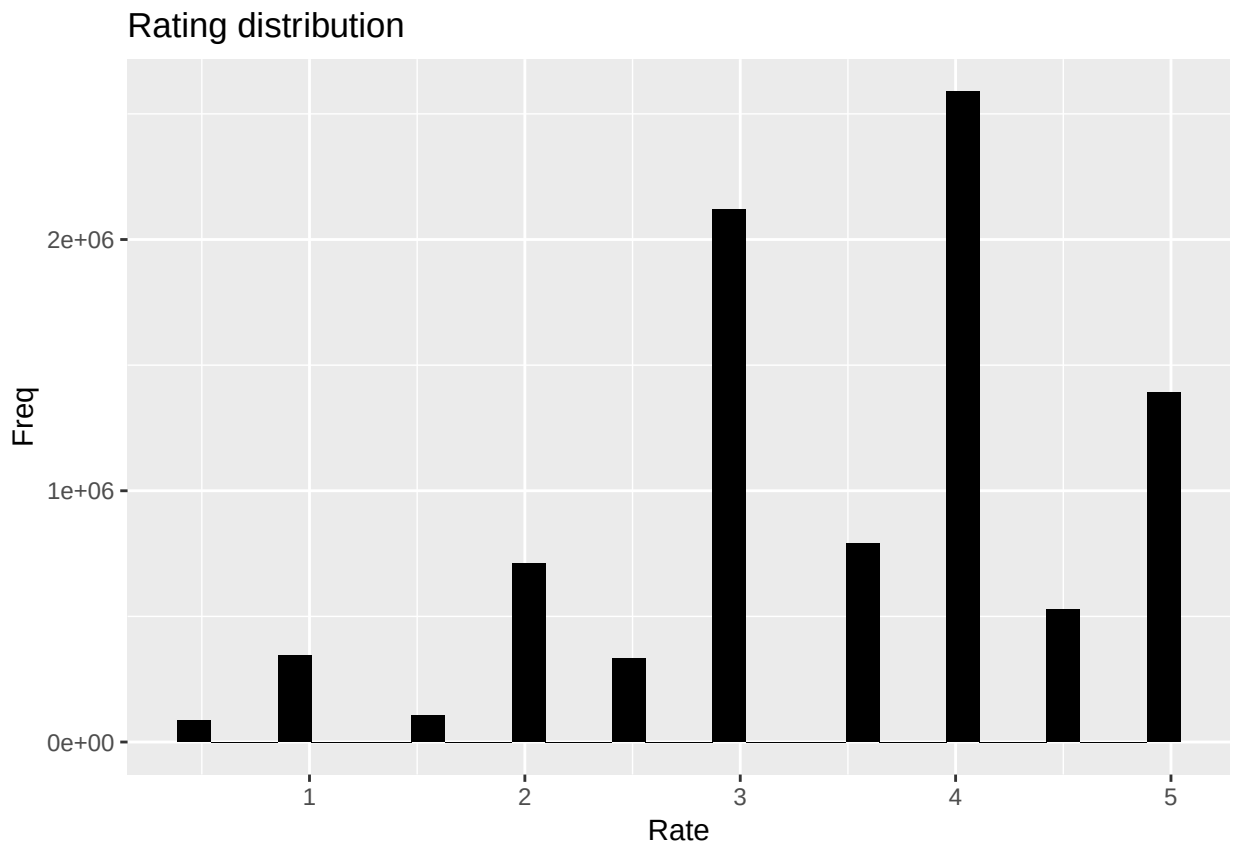
2.1 Exploratory data analysis

The data set contains 9000055 observations of 6 variables.

- `userId`: Unique identification number given to each user.
- `movieId`: Unique identification number given to each movie.
- `timestamp`: Code that contains date and time in which the rating was given by the user to a specific movie.
- `title`: Title of the movie.
- `genres`: Motion-picture category associated to the film.
- `rating`: Rating given by the user to the movie. From 0 to 5 stars in steps of 0.5.

The main goal is to create a model capable of predicting the variable rating. As we can see by the data below the average score is near to 3.5 with positive value more common than negatives.

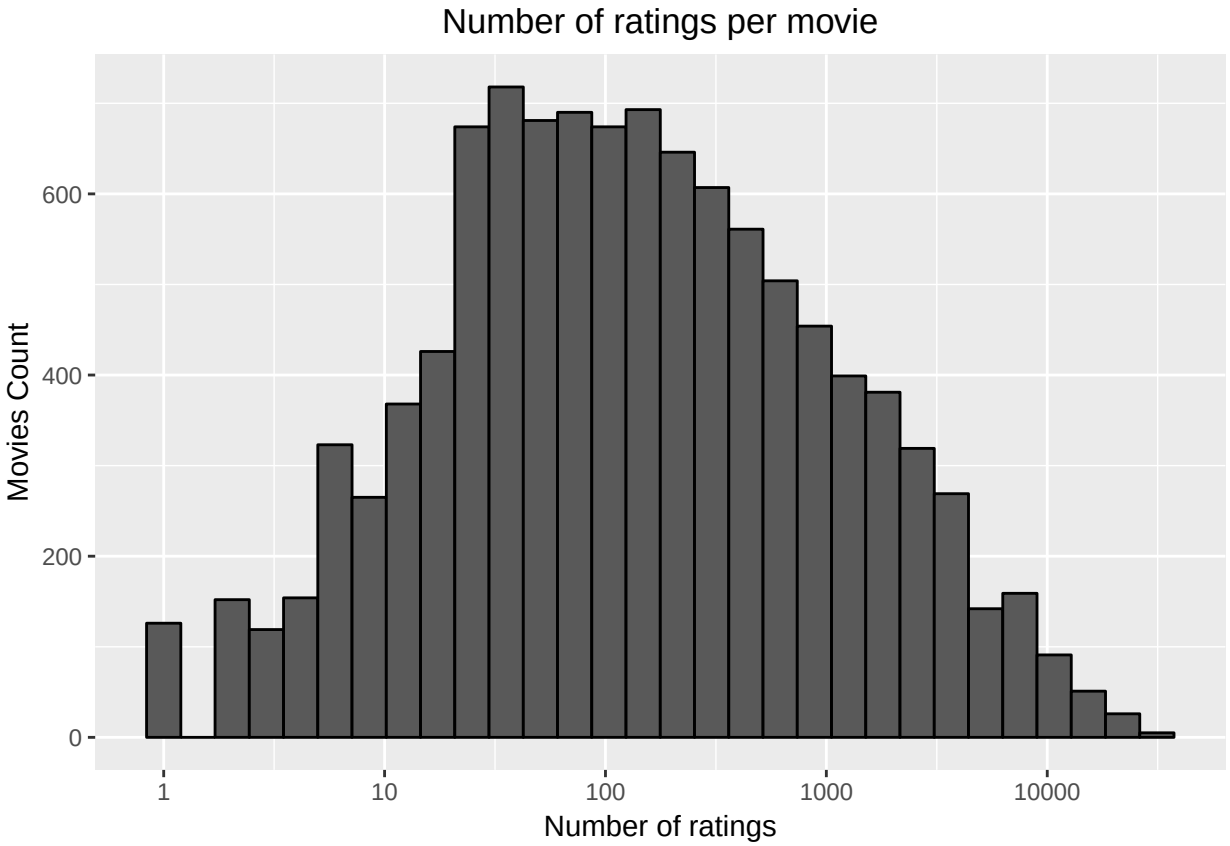
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.500   3.000   4.000   3.512   4.000   5.000
```



The data contains more users than movies:

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

Some movies get rated more than others:



2.2 Model creation

The RMSE is defined as:

$$RMSE = \sqrt{\frac{1}{N}\sum_{u,i}(\hat{y}_{u,i} - y_{u,i})^2}$$

where $y_{u,i}$ are defined as the rating for movie i by user u and denote our prediction with $\hat{y}_{u,i}$ with N being the number of user/movie combinations and the sum occurring over all these combinations.

We will use different models and then choose the one that minimize the RMSE.

First of all we will split the edx dataset in train and test set:

2.2.1 Basic model

We start with the simplest possible recommendation system: we predict the same rating for all movies regardless of user. This prediction can be found using a model based approach. A model that assumes the same rating for all movies and users with all the differences explained by random variation would look like this:

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

with $\epsilon_{u,i}$ independent errors sampled from the same distribution centered at 0 and μ the "true" rating for all movies. We know that the estimate that minimises the RMSE is the least squares estimate of μ and, in this case, is the average of all ratings

```
## [1] 1.061202
```

Whith a RMSE of 1.06 on a scale 0.5 to 5 this model is almost useless.

2.2.2 User and Movie effect Model

We know from experience that some movies are just generally rated higher than others. This intuition, that different movies are rated differently, is confirmed by data. We can augment our previous model by adding the term b_i to represent average ranking for movie i: and b_u as a user-specific effect.

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

The b's are referred to as effects.

We estimate this effect by computing μ and estimating b_i, as the average of

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

First we evaluate the new model on the test edx dataset:

```
## [1] 0.8435874
```

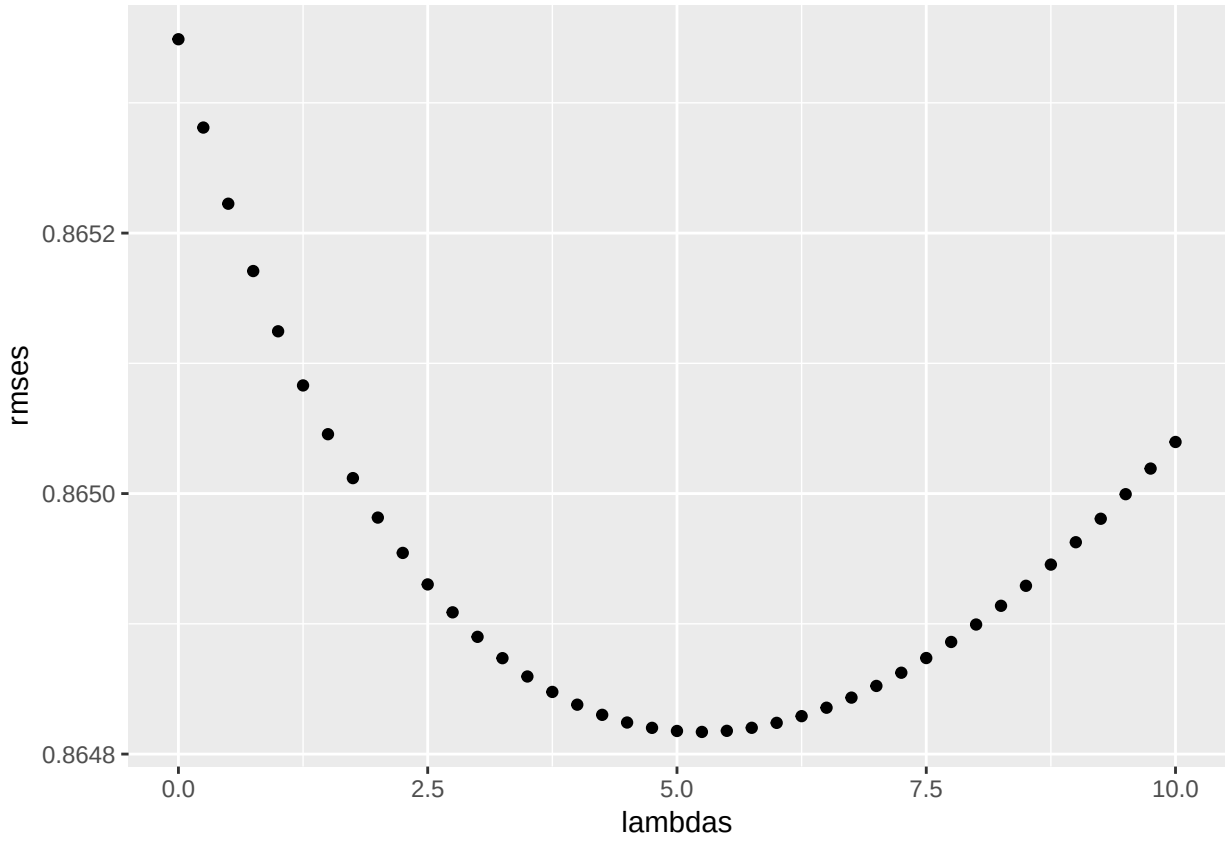
Now we can evaluate the prediction model on the validation dataset

```
## [1] 0.8821153
```

We can notice that the RMSE on the validation dataset is higher.

2.2.3 Regularized user and Movie effect Model

Regularization is a form of regression, that constrains or shrinks the coefficient estimates towards zero. In other words, this technique discourages learning a more complex or flexible model, so as to avoid the risk of overfitting. Here we use the concept of regularization in order to account for the effect of low numbers of ratings both for movies and users. The regularisation process will evaluate different values for λ, delivering to us the corresponding RMSE.



```
## [1] 5.25
```

```
## [1] 0.864817
```

3. Results

The table below summarises the RMSE's of the developed models:

```
## # A tibble: 3 x 2
##   Method                                  RMSE
##   <chr>                                  <dbl>
## 1 Base                                   1.06
## 2 User and Movie Effect on validation    0.882
## 3 Regularized Movie and User Effect Model 0.865
```

4. Conclusions

We can interpret the RMSE similarly to a standard deviation: it is the typical error we make when predicting a movie rating. The resulting data shows us that the Regularized Movie and User Effect Model has the lowest RMSE and is the best fitting model. A RMSE equal to 0.8648 can be considered a good score for a movie recommandation system.