MovieLens

Introduction

This document describes the main steps taken to develop a model predicting the variable `rating` in the test set `validation`, after being trained with the information contained in the training set `edx`. Both sets represent a small sample from a comprehensive data source about MovieLens, an online movie recommendation system.

Out of the five predictors included in the input datasets, the current model is only considering `userId` (integer) and `movieId` (integer). Some of the remaining variables might contain relevant information and their analysis is a matter of future work.

The model is based on the following main ideas:

- Under these conditions, averages provide good preliminary estimates.
- The data is irregularly distributed (e.g., some movies get many more ratings than others).
- The predictions are the result of adding the effects at three different levels: all the movies, each individual movie and each individual user.

The accuracy of the generated predictions is determined using the root mean square error (RMSE). The obtained value (0.86471) indicates a good performance.

Analysis

Firstly, it is important to adequately understand the underlying reality with a major focus on the variable to be predicted. It can be summarised in these points:

- Subjective assessments tend to be somehow conditioned by the given environment. That is, the ratings issued by the users in MovieLens are likely to meet certain common criteria.
- A significant proportion of movies tends to be liked or disliked by most of people. Scenarios where opinions are distributed approximately equally can be considered almost exceptional.
- Usually, somewhat homogeneous groups of people like and dislike different types of movies.

The following graphs show how ratings are distributed across a group of fifty random users and movies. It is clear that some movies are rated more often than others, and also that some users are more active contributors. Both conclusions are fully compatible with what is expected, but the effects from these irregular distributions of ratings need to be somehow addressed.

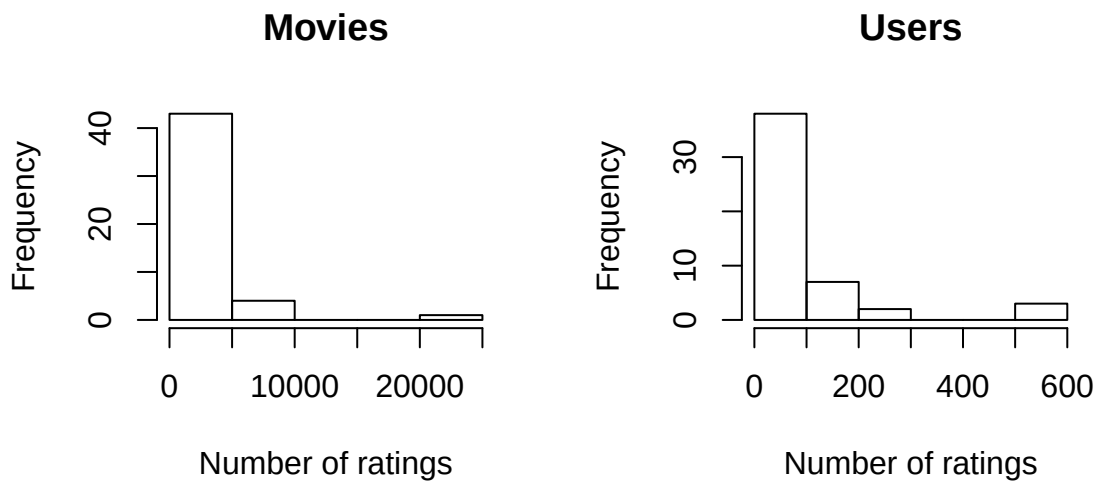

Figure 1: Rating frequencies.

By applying the previous ideas, a preliminary modelling approach can be defined with the following equation:

$$rating_{u,m} = aver_r + \epsilon_{u,m}$$

It outputs the rating given by each user to each movie, with $\epsilon_{u,m}$ representing the independent errors and $aver_r$ the average value of all the ratings across the whole data set. Note that this last variable is already accounting for the aforementioned environmental effects via assuming that all the ratings are somehow related.

As expected, the accuracy of such a simplistic model is not particularly good. More specifically, the proposed loss function (i.e., RMSE) delivers a modest 1.0612. In any case, it does already represent an acceptable starting point. Additionally, its average-based essence also seems the best way to go further.

Some movies tend to be liked more than others and this fact has to be taken into account when estimating user ratings. This preliminary model can, consequently, be improved by bringing all those movie-specific effects into consideration. Mathematically, it can be represented as the average difference between the ratings for the given movie and the default estimates (i.e., predictions generated by the model as defined up to this point). The term "bias" adequately describes this and subsequent additions. The new model is hence defined by:

$$rating_{u,m} = aver_r + bias_m + \epsilon_{u,m}$$

There is still a relevant aspect which has been ignored so far. Users also have their own preferences and tend to like some movies more than what most people do. This can again be modelled as the average of the variations between the predictions so far and the ratings issued by the given user to all the movies. After this last inclusion, the model becomes:

$$rating_{u,m} = aver_r + bias_m + bias_u + \epsilon_{u,m}$$

The previous extensions have improved the accuracy of the model until getting a RMSE value of 0.86535. This is appreciably better, but it can still be improved further.

The uneven distribution of ratings across movies and users is likely to have an appreciable impact on the reliability of the model. Regularisation seems the best way around this problem: predictions generated from small sample sizes will be penalised. Mathematically, the model is updated by redefining the way in which averages are being calculated via adding a constant (`lambda`) to the corresponding total number of elements in the denominators.

The best value for the aforementioned `lambda` parameter is intrinsically associated with the given data and its determination is not immediate. Logically, only the information contained in the training data set can be used to find that value out. On the other hand, it is definitely possible to perform some modifications in order to better resemble the actual testing conditions. More specifically, the training data set can be divided into two subsets (training and testing) to ensure as realistic conditions as possible. That is, various candidates (i.e., the model up to this point including different values for the new `lambda` constant) are trained on and the accuracy of their predictions compared against the corresponding data subsets. The best `lambda` (4.5) is assumed to be the one from the candidate model with the smallest RMSE.

As shown in the table below, the top and bottom predictions are outside the training data boundaries (i.e., 5 and 0.5). These outliers surely have a negative impact on the RMSE value. The correction is trivial: replacing these outliers with the aforementioned boundaries.

Table 1: Top and bottom predictions.

Top	Bottom
6.00400	-0.41261
5.89233	-0.36965
5.81080	-0.27134
5.77304	-0.18507
5.76040	-0.11831
5.69677	-0.05657
5.69015	-0.04146
5.67709	-0.03997
5.67161	-0.03569
5.66602	-0.02093

Results

By bearing in mind that, under the current conditions, a RMSE value below 0.86490 can be considered a success, the performance of the final model (0.86471) is undoubtedly good.

The model predictions are showing a strong correlation (0.80849) with respect to the original ratings when considering averages across all the movies.

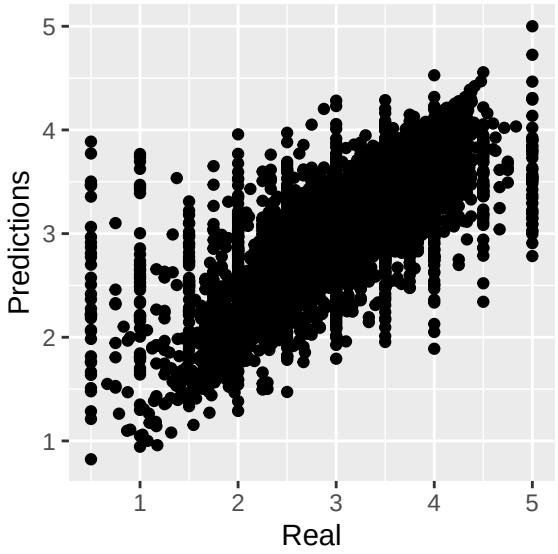

Figure 2: Correlation of movies.

When grouping by users, the correlation between the predictions and the original values is a bit weaker (0.69773), but still good enough, as shown in the graph below.

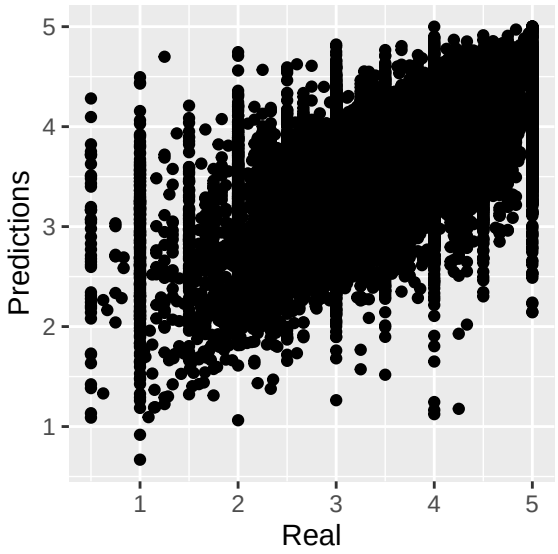

Figure 3: Correlation of users.

Additionally, the generated predictions will always lie within the training set boundaries and, consequently, no unrealistic ratings are expected.

Conclusion

The created model is undoubtedly fulfilling the main goals and expectations which motivated this work. Its accuracy is appreciably better than the most optimistic target. The delivered predictions are realistic and closely resemble the actual values under a wide variety of different scenarios. This model can definitively be assumed to somehow understand the reality underlying the input data sets.

The previous paragraph should not hide the fact that some improvements are certainly possible. At the user level, for example, the performance of the model is a bit lacking. It does not seem to account for the user-to-user variability as well as it could.

The current model is not even considering some variables which might also contribute to providing a better picture. More specifically, analysing the effect of movie genres, especially on how users rate movies, is a matter of future work.

In summary, the model described in this paper can be considered a solid first step to face the numerical understanding of the data under consideration and, by extension, of movie ratings from Movielens or from any other source. Further work is certainly expected, but the main ideas exposed here are likely to remain essentially unaltered.