Project Report MovieLens

Samuel Kuenti

17. April 2022

Introduction

The task for this first project of the Capstone module was to predict movie ratings based on a machine learning algorithm.

The data came from the publicly available MovieLens-databse. The data were split into a training set, a test set and the validation set. The latter was to be used only after the completion of the model in order to assess its performance (model performance was assessed by comparing RMSE values, as described below).

Here, as a first step, the movie and user effects described in the course were modeled. Then, it was determined that there was a small genre effect in the data. This effect was then added to the model in order to improve the overall performance.

Finally, this refined model was assessed by means of the validation set of the original data, which mimicked model performance on actual 'real-world' data.

Methods

Data preparation

The data were downloaded from the grouplens homepage (http://files.grouplens.org/datasets/movielens/ml-10m.zip).

Then, 10% of the whole dataset were set aside as the final validation set, as per the module requirements. This holdout data was to be used to assess the final model and to allow objective comparison of this particular solution to other student's algorithms.

The modeling data was then split further into a training set (90%) and a test set (10%). This splitting was chosen because the model's structure was clear from the beginning, and no tweaking was expected in response to model performance. Therefore, it was appropriate to have most of the data for the tuning of the model.

The training set consisted of 8100048 cases.

Model performance had to be assessed by calculating the root mean squared error (RMSE), defined as follows:

$$RMSE = \sqrt{\frac{1}{N} \cdot \sum (\hat{y} - y)^2}$$

The lower the RMSE, the better the model's predictions.

Initial model from the course

Initially, a simple model from the course was recreated. This consisted of using the overall mean rating of all the movies as a starting point ($\mu = 3.51246$).

Then, a movie bias

$$\hat{b}_m = \frac{1}{N} \cdot (y_m - \hat{\mu})$$

and a user bias

$$\hat{b}_u = \frac{1}{N} \cdot (y_{m,u} - \hat{\mu} - \hat{b}_m)$$

were estimated, and the inital model looked as follows:

$$\hat{y}_{m,u} = \hat{\mu} + \hat{b}_m + \hat{b}_u$$

This initial model achieved an RMSE of 0.86469 on the training set.

Genre effect

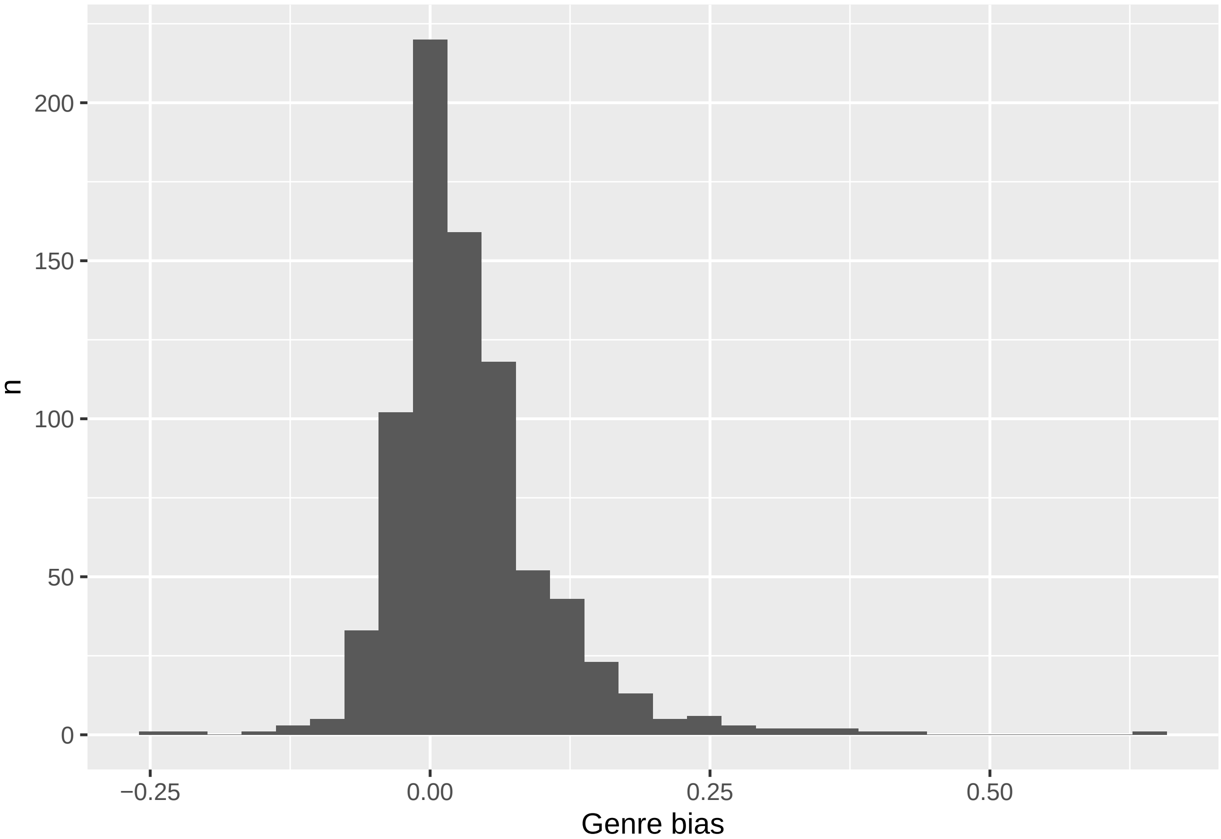
In the textbook, the idea of a genre effect was brought up (Chapter 33.8). The genre column in the data includes every genre applying to a movie. For simplicity's sake, every combination of genres was treated as a distinct level of the genre factor.

A genre bias could then be defined as follows (with g indexing genre combinations):

$$\hat{b}_g = \frac{1}{N} \cdot (y_{m,u,g} - \hat{\mu} - \hat{b}_m - \hat{b}_u)$$

In the training set, there were 797 genre combinations.

A histogram revealed small genre effect:



Not surprisingly, looking at the standard deviations of the biases showed that the additional variation accounted for by the genre effect was rather small, and that therefore model performance could not be expected to improve much.

Table 1: SD for different biases

bias	SD
Movie	0.57194
User	0.41385
Genre	0.07306

Nevertheless, refining the initial model by this genre effect seems viable.

Assessment of the refined model

The genre bias was added to the model:

$$\hat{y}_{m,u,g} = \hat{\mu} + \hat{b}_m + \hat{b}_u + \hat{b}_g$$

The inclusion of the genre bias improved the model performance slightly to an RMSE of 0.86433.

It must be noted that the addition of the genre bias improved the model performance by a mere 0.04167 percent.

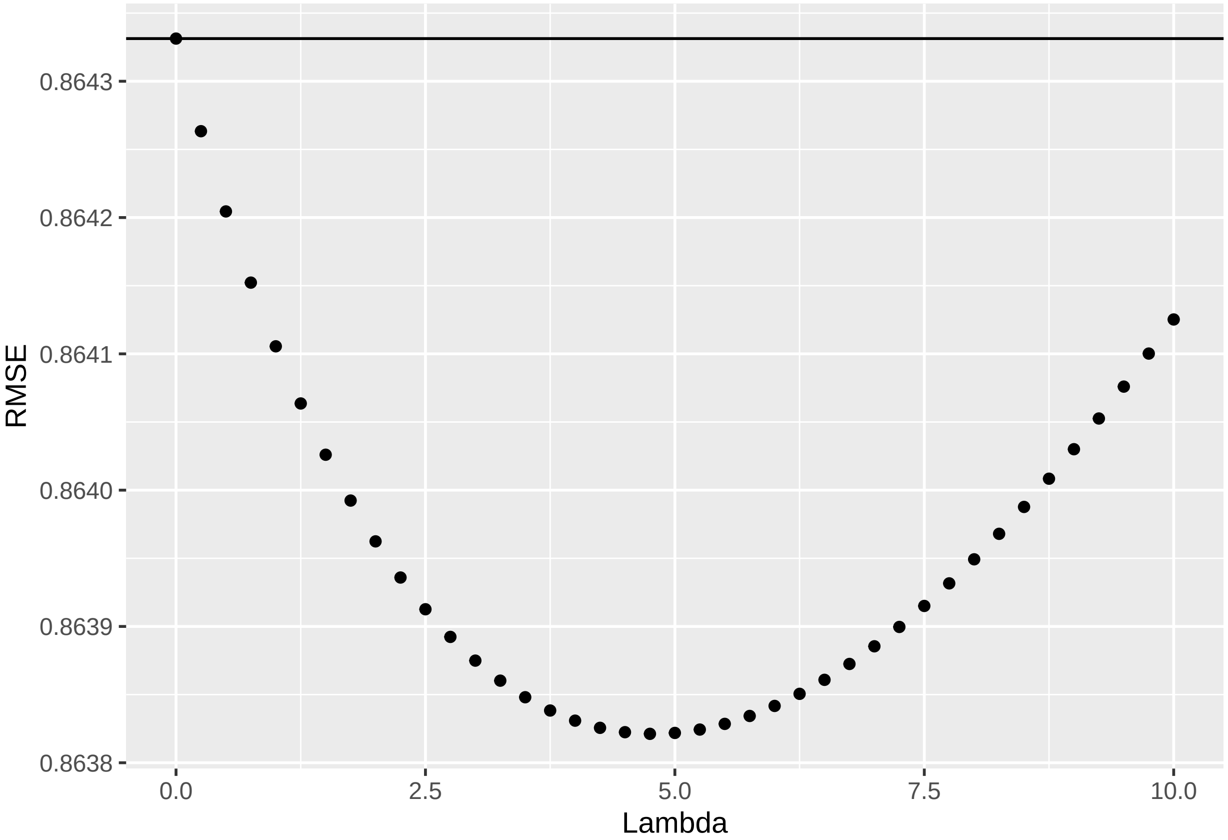
Adding regularisation of the bias terms

From the above, it seemed clear that the model needed further refinement. One approach described in the course was regularisation, with the aim of penalising large estimates coming from small sample sizes.

In order to achieve this, a tuning parameter λ was cross-validated to minimise the total error term:

$$\frac{1}{N} \cdot \sum_{m,u,g} (y_{u,i,g} - \mu - b_m - b_u - b_g)^2 + \lambda(\sum_m b_m^2 + \sum_u b_u^2 + \sum_g b_g^2)$$

Plotting RMSEs for different λ showed that this lead to an improvement of the model. The horizontal line in the plot shows the RMSE of the refined model, prior to regularisation.



A λ of 4.75 minimised the RMSE at 0.86382 and was therefore included in the final model.

Results

The final validation set was used to assess model performance.

The refined model achieved an RMSE of 0.86545, just reaching the 15 point range on the grading scale.

The regularised model achieved an improved RMSE of 0.86486 when used on the final validation set, just below the target RMSE of 0.86490 for the 25 point score.

Conclusion

The task in this project was to develop an algorithm to predict movie ratings. Here, a simple initial model from the course was refined by the inclusion of a third error term, based on genre effects.

The goal of improving the initial model by adding an additional bias term was achieved. However, the model initially improved just slightly, and it must be noted that the model improvement after the addition of the genre bias was so small that it could have been a random result.

As a result, regularisation was implemented to refine the model. This led to a further improvement, with the final model reaching an RMSE below the threshold for the maximum model performance score.

Training the model on 90% of the training data could have led to overfitting. As a next step, the issue of overfitting could be investigated and model performance could potentially be improved further as a result.