MovieLens Capstone Project

Esther Sienkiewicz

02/12/2021

Introduction

In this report, the objective is to predict movie ratings in MovieLens dataset using machine learning (ML) algorithms. The dataset includes the following fields: User, Movie, Movie Title, Timestamp and Genre. The analysis explores different impacts the corresponding data has towards accurately predicting the movie rating. For example, some obscure movie titles (i.e. non blockbuster, independent films) may contain few or even a single user rating. This report will explore the effects this has on the outcome of the predicted movie rating.

The work herein is the accumulation of the learning from PH125.1x to PH125.8x Data Science program.

Evecutive Summary

The predicted results will then be evaluated against the residual mean squared error (RMSE). For this case we define the rating for the movie i by user u and denote the prediction $\hat{y}$ and y with N as the total number of user/movie combinations.

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

It can then be said that the RMSE is the figure of merit to which the ML aglorithm will be judged upon.

Analysis

The first step is to develop a baseline model to understand what the basic RMSE of all the moving ratings within the MovieLens dataset. From here, the impact of each additional method or algorithm applied can be evaluated its effectiveness. Therefore, the difference of each rating away from the real rating can be represented as $\varepsilon_{u,i}$ in the expression below:

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

In other words, the error ($\varepsilon_{u,i}$) is the value in which the goal is to minimize. The average of all ratings is calculated by:

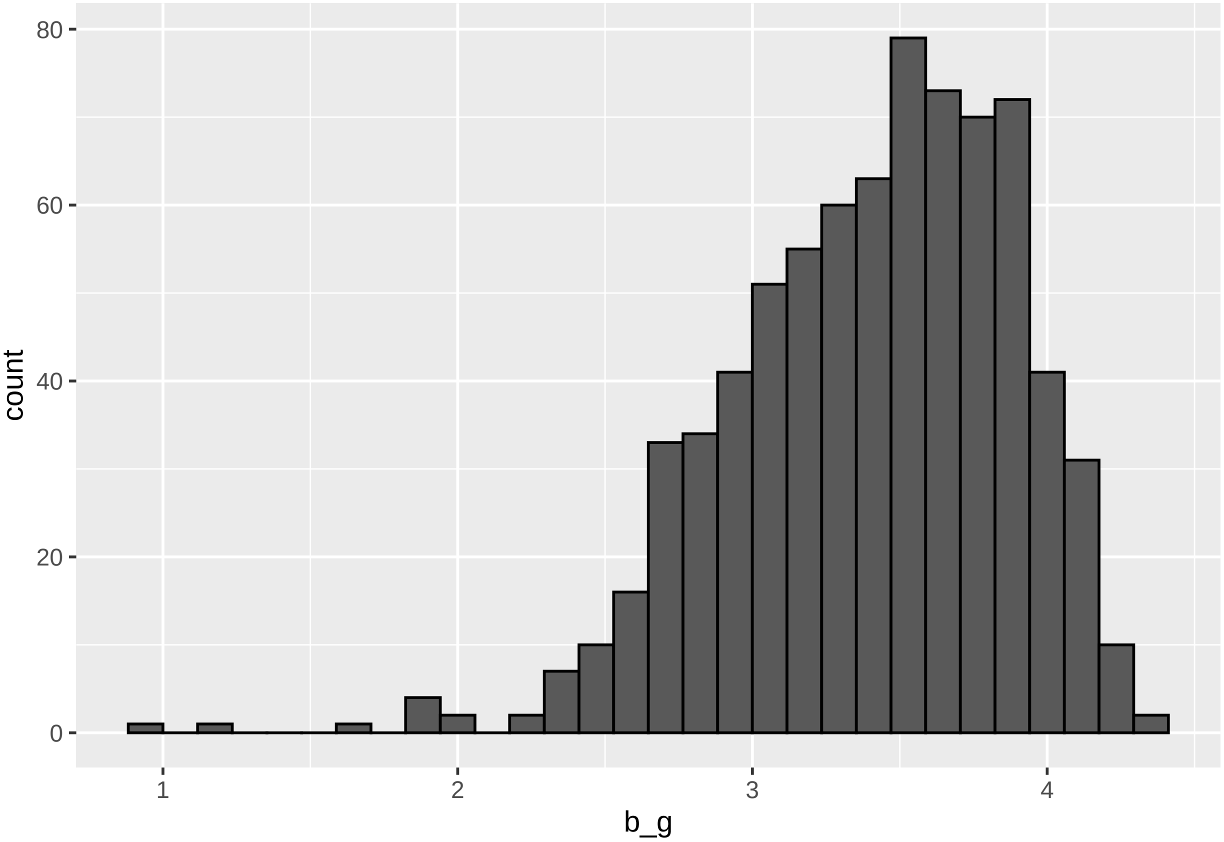
```
Y <- mean(edx_train$rating)
RMSE(edx_test$rating, Y)
```

```
## [1] 1.060025
```

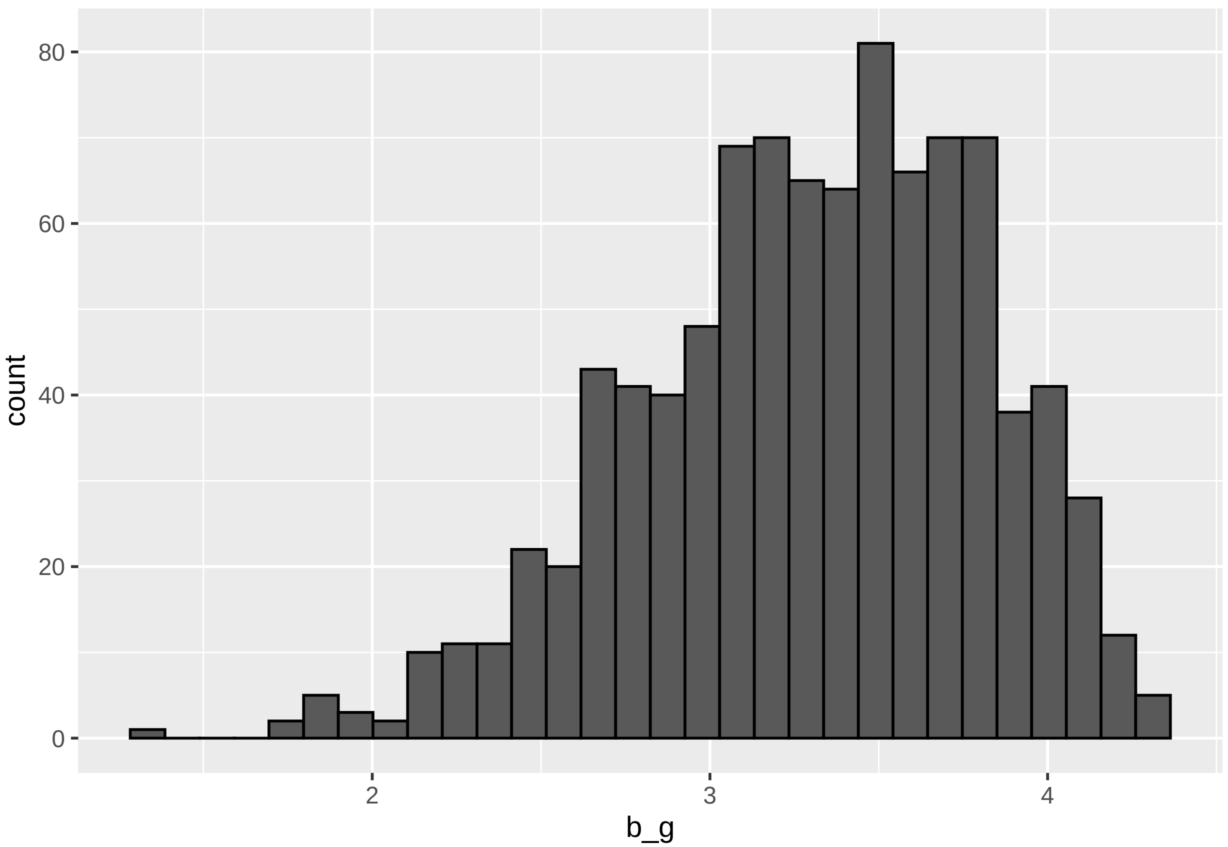
Using the RMSE equation stated in previous section, comparing the validation set with the current edx_train set we have an RMSE of 1.06.

There are 3 bias which will be examined as part of this report: movie effects, user effects, and genre effects.

First the movie and user effects applied only help to reduce the RMSE down to 0.869. It helps by 0.191 however with some penalization of certain ratings it could help further reduce the RMSE. Next, we look also look at how the effect of genre can help. We looked at different genres and Thrillers is one of the main genres where there is large user rating variability. One hypothesis on this is probably due to the fact that how each person perceives a good thriller (too scary or not enough scary) varies widely unlike Romance genre where usually predictable endings of happy ever after. In the first plot (below) we see that the variance is not large as many users are rating romance movies do not seem to vary off from 3-4 rating.



Next we look at the movies which are classified as a Thriller genre and there is a slight increase in variation between 2-4 rating.



We examined the genre aspects and still the RMSE does not seem too much effected, next we examine the penalization methodology explained just above. In the course material this was referred to as "Regularization". In the least square model we showed above we add the b_i variable which was intended to estimate the movie effect shown above.

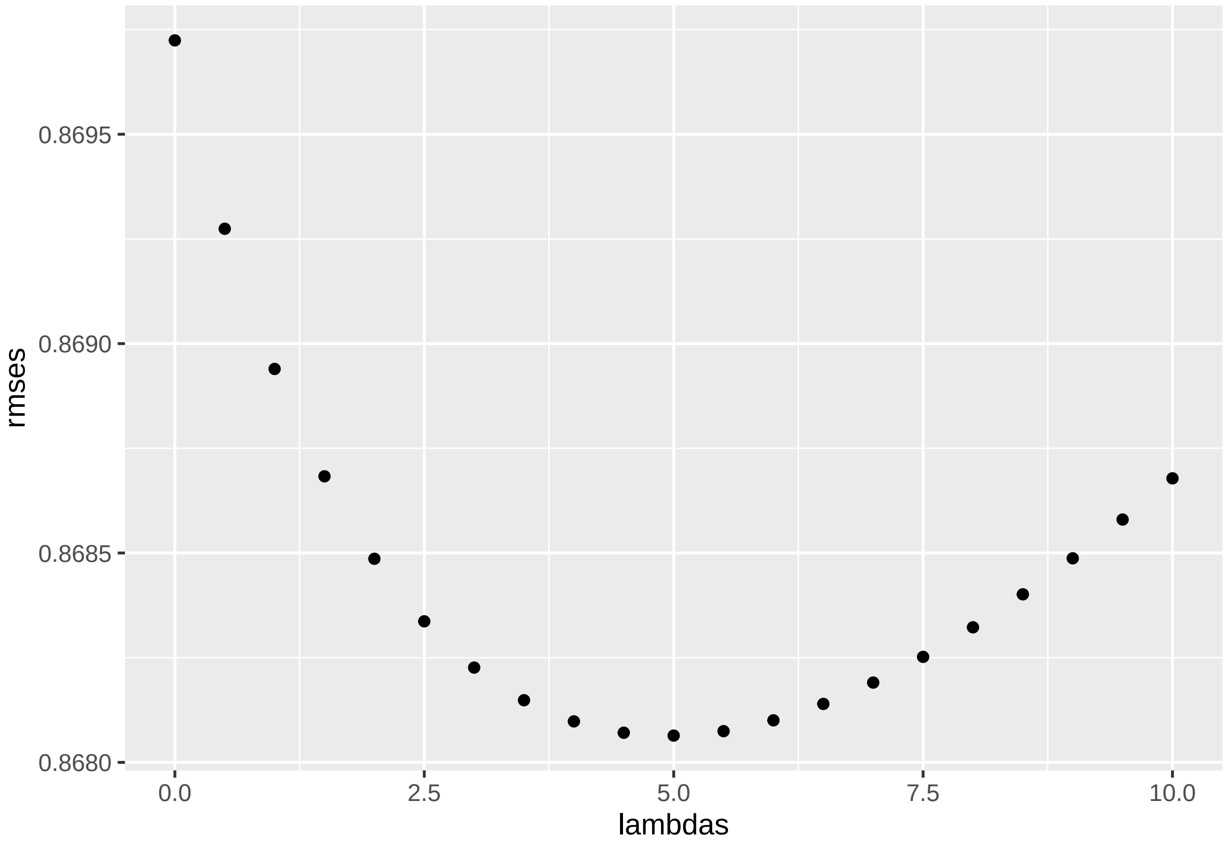
$$Y_{u,i} = \mu + b_i + \varepsilon_{u,i}$$

Further to model the penalization will be applied upon each total variability of the movie effects by λ

$$\lambda \sum_i b_i^2$$

where λ is the penalization factor. We first need to know what value we should assign λ

```
qplot(lambdas, rmses)
```



```
lambdas[which.min(rmses)]
```

```
## [1] 5
```

Then if we plug back λ as 5 into the penalization factor we hope to see a reduction of the final RMSE. The penalization did not help as much as predicted, it only reduced the RMSE by 0.0002731, but still helpful in the right direction.

Results

The evaluation of each modelling can be broken down into the follow:

Method	RMSE
Just the mean	1.060025
Movie Effect Model	0.9440085
Movie and User Effect Model	0.8697242
Movie, User and Genre Model	0.8693637
Regularized Movie Effect Model	0.8694511

As it can be seen in the table above, if the regularization method has been applied to more than the movie effects model the final RMSE may be in fact lower. However, the work shows that taking all factors of the models (movie, user, genre) effects prior to regularization would help.

Conclusion

It would have been interesting to explore the Genre field in more details. Does a user whom generally watches one genre (i.e. majority of movies rated is horror) rate other movies such as dramas more critically? We tend to favor certain types of movies and perhaps have bias. This would be an interesting place to apply some further regularization method.