MovieLens Recommender System Capstone Project - Report

Pedro Bernardino Alves Moreira

28 Feb, 2020

Contents

1 Introduction

This project aims to create a recommender system using the MovieLens dataset as part of the final grade for the **HarvardX PH125.9x Data Science Capstone** on edx.

It uses only a subset of the movielens, containing approximately 10 millions movies ratings, being 90% used for training and 10% for validation. The project gradually evolves an algorithm for recommendation, starting with the simplest possible model then adding more features and comparing to the previous models. Are made a total of 6 models, from just the average to a regularized movie and user effect.

The recommender system works by trying to predict what rating a user would give for a random movie, these movie rating predictions will be compared to the true ratings in the validation set using the root mean squared error (RMSE), with the following formula:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n} e_t^2}$$

The final model **Regularized movie and user effect with indenpendent tuning parameter** obtains a result of 0.8566827 RMSE.

2 Methods

2.1 Naive Model - Just the average

The naive model is a separator line between the guessing and the real analysis. This means that it is the simplest model someone can build, any guessing rating should have a worst RMSE just as any other serious model should have a better RMSE.

In this case, the simplest model is the mean for all ratings, so we are predicting that all ratings a user probably will give, are somehow towards the mean of all ratings, that is **3.5124652**. The formula used is:

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

The $\hat{\mu}$ is the mean from all ratings in the training set, the "true" rating. And the $\varepsilon_{i,u}$ is the independent errors sampled from the same distribution centered at 0.

Table 1: Naive model results

	RMSE	Grade
Just the average	1.061202	5

If we run the same model with any value different from the mean, the RMSE is worst, as we can see in table 2.

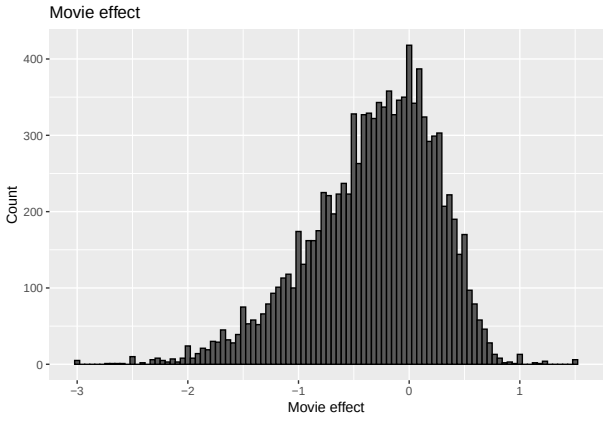
Table 2: Random values naive model

	RMSE	Grade
Random value lesser than the average	1.519737	5
Random value greater than the average	1.402808	5

2.2 Movie effect model

The naive model can be improved to get a better RMSE. The first approach is to add the movie effect. Some movies are generally ranked higher than others, probably because they are better movies.

We can see it if plot the density for the distance between each movie rating mean and the all movies rating mean:



The movie effect model can be defined as:

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

It is the previous model with the new movie effect parameter. The b_i measures the popularity for each

movie i. Where b_i is the difference from each movie rating mean to all movie rating mean:

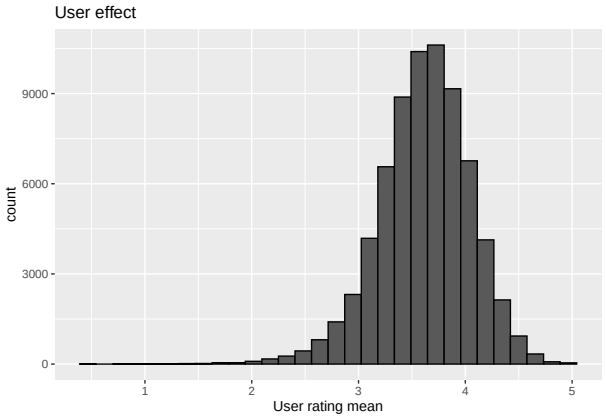
$$b_i = |\hat{\mu}_i - \hat{\mu}|$$

Table 3: Movie effect results

	RMSE	Grade
Movie effect	0.9439087	5

2.3 Movie and user effect model

Then, we also add the user effect. Some users tend to give higher rating than others. We can see this by plotting the average ratings by users who have rated more than 100 movies:



Now the model is defined as:

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

And the new parameter b_u is the difference from the user rating to the all movie rating mean with the movie effect:

$$b_u = |\hat{\mu}_u - \hat{\mu} - \hat{b}_i|$$

Table 4: User and movie effect results

	RMSE	Grade
User and movie effect	0.8653488	15

2.4 Regularized movie effect model

There are movies rated many times just as there are movies rated only few times, sometimes only once. Since the movie effect relies on the quantity of times a movie has been rated, the more rates the movie get, the more precise the movie effect is. So we need to regularize this by putting more weight when a movie is rated many times, and shrinking the movie effect when a movie haven't many reviews.

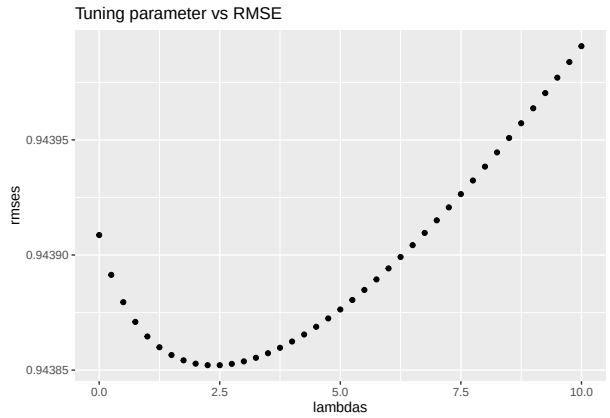
We can see this ploting how many times each movie was rated and the movie effect for that movie, the red line separate between movies with more or less than 10 reviews:

To regularize the movie effect we will use the following equation:

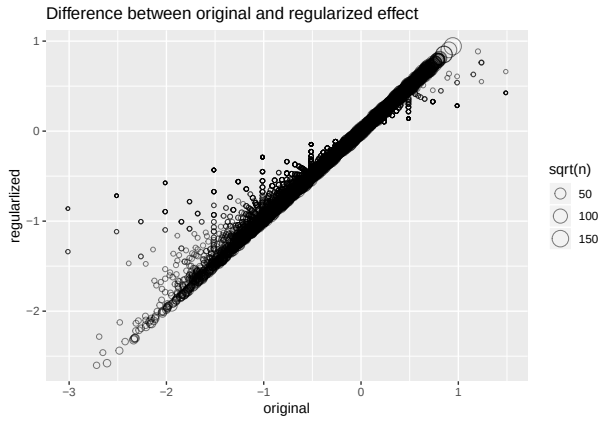
$$\hat{b_i}(\lambda) = \frac{1}{\lambda + n_i} \sum_{u=1}^{n_i} (Y_{u,i} - \hat{\mu})$$

Where λ is the tuning parameter that will optimize the formula, n_i is the number of rating made for movie i. The larger the tuning parameter λ is, more we shrink when the number of ratings is small.

Here is a cross-validation to find the best value for λ that minimizes the RMSE:



Then, with the best tuning parameter we can see how the not regularized and the regularized movie effect are differents:



The regularized movie effect has a worst RMSE than the combined not regularized movie and user effect. However we see a very small improvement from the not regularized movie effect.

Table 5: Regularized movie effect results

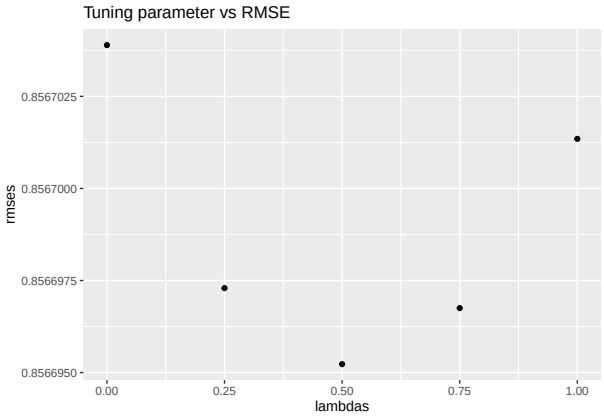
	RMSE	Grade
Regularized movie effect	0.9438521	5

2.5 Regularized movie and user effect - same tuning parameter

Just as the regularization was applied to the movie effect, this model has the regularization applied to both movie and user effect. Similar to the previous, the penalty equation is the following:

$$\hat{b_u}(\lambda) = \frac{1}{\lambda + n_i} \sum_{u=1}^{n_i} (Y_{u,i} - \hat{\mu} - \hat{b_i})$$

Here is the plot to find the best tuning parameter for both user and movie effect:



The regularized movie and user effect has a better RMSE:

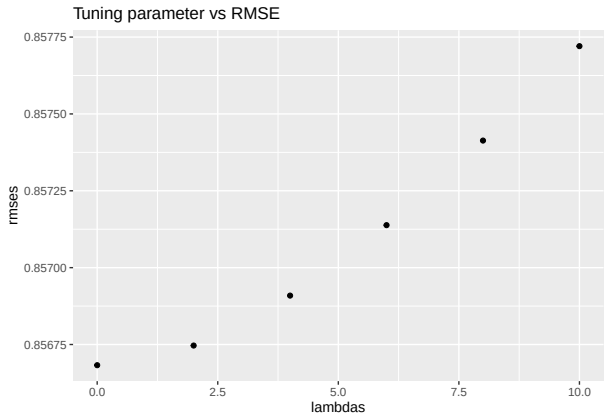
Table 6: Regularized movie and user effect with same tuning parameter results

	RMSE	Grade
Regularized movie and user effect same tuning parameter	0.8566952	25

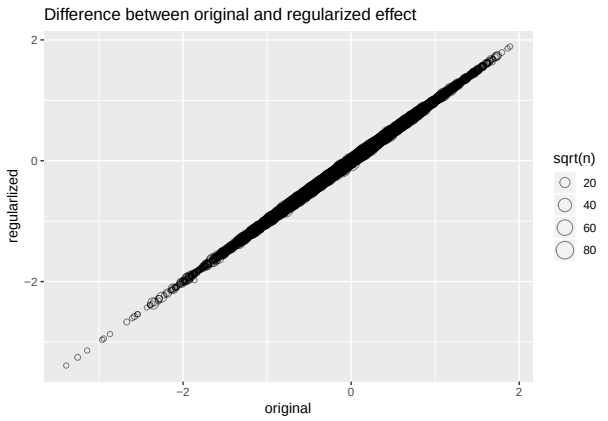
2.6 Regularized movie and user effect - independent tuning parameter

Instead of using the same tuning parameter for both user and movie effect, we can use the best for each.

Here is the plot to find the best tuning parameter just for the user effect:



With the best tuning parameter we can see how the not regularized and the regularized user effect are differents, and we actually see there were little difference:



The regularized movie and user effect with independent tuning parameter has best RMSE:

Table 7: Regularized movie and user effect with independent tuning parameter results

	RMSE	Grade
Regularized movie and user effect	0.8566827	25

3 Results

After gradually evolving the algorithm to reduce the RMSE, it's clear that both user and movie effects have a high influence on the rating. The best model "Regularized movie and user effect independent tuning parameter" has a RMSE of 0.85668.

Here is the final table with all results:

Table 8: Final results

	RMSE	Grade
Just the average	1.0612018	5
Movie effect	0.9439087	5
User and movie effect	0.8653488	15
Regularized movie effect	0.9438521	5
Regularized movie and user effect same tuning parameter	0.8566952	25
Regularized movie and user effect	0.8566827	25

4 Conclusion

Further investigation would be exploring the genre effect relationship with user and movie effects. Perhaps a better approach would be using the Matrix Factorization technique and Principal Component Analysis with the recommenderlab package.