Movie Rating Prediction with Regularized Linear Regression Model (MovieLens Project)

Wan Chi Fung,Victor

2020/8/28

1. Introduction

Rating prediction is important for movie recommendation system on item specific recommendation. Different factors including movie effects, user effect, movie released year and genre of the movie could affect the movie rating. The projects aims to predict movie rating on a given movie data set (movielen). In the following report, a model is constructed to predict rating though considering different effects on the movie rating. The loss function, residual mean squared error (RMSE), is used for evaluation.

2. Methods

2.1 Dataset Preperation

The data set is provided by the GroupLens research group. It contains userID, movieID, rating, timestamp, title and genres of the movie.

After downloading the data set in the r-script, the data set is then seperated into 2 sets of data, edx and validation, via the caret package. For the edx data set, it is used for model train and testing. For the validation, it is the validation set for model evaluation. The ratio of numbers of data in edx and validation is 9:1.

2.2 Data Wrangling

Data wrangling is performed on the edx data set which includes extracting released years from the title column and converting datetime format on the timestamp column.

2.3 Create train set and test set

Train set and test set are create though seperating the edx data set. The ratio of train set and test set is 8:2.

2.4 Evaluation matrics

For model evaluation, a typical loss funtion, Resdiaul Mean Square Error (RMSE) is used. And it is defined as:

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

where N is the number of user/movie/year/genre combinations and the sum of all these combinations.

If the RMSE is larger than one, it means the typical error of the model is larger than one star, which is an undesire result. Therefore, the model should keep the RMSE below one and reduce it as small as possible.

2.5 Model Building

First of all, the mean of all movie's rating is considered to assume the same rating for all movies and users with all differences explained by random variation:

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

where $\varepsilon_{i,u,y,g}$ is the independent errors sampled from the same distribution and the μ is the averaged rating of all movies.

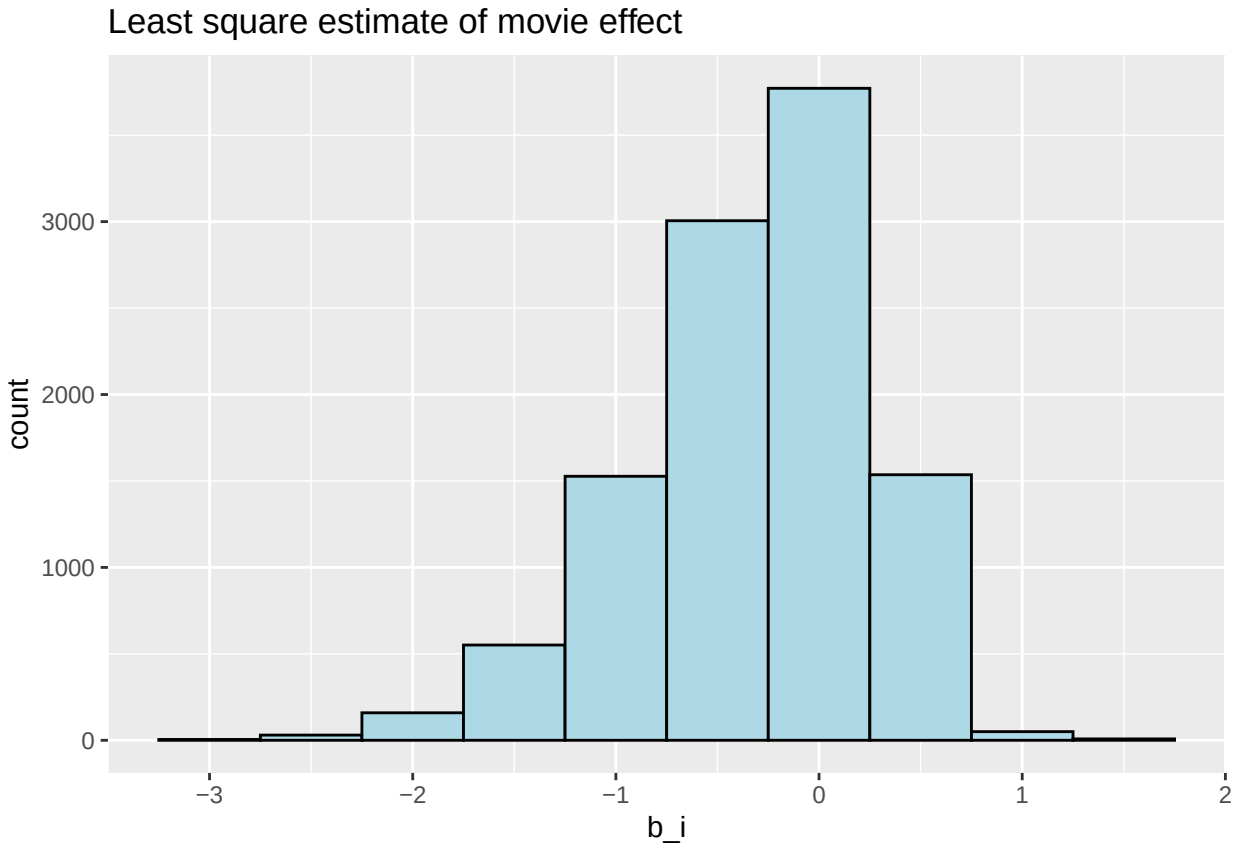
```
## The average rating of movie in train set is 3.512574
```

method	RMSE
Average rating	1.060704

The RMSE of just using average rating as predictor is 1.060704.

2.5.1 Movie Effect

Some movies maybe rated generally higher than some other movies. To confirm this, the least square estimate of movies are plotted in the following:



From the above plot, the estimate of movie effect vary substantially. Some movies are rated 1 or 2 stars lower than the average and some are rated 1 stars higher than the average.A movie bias term, b_i should be added to the model:

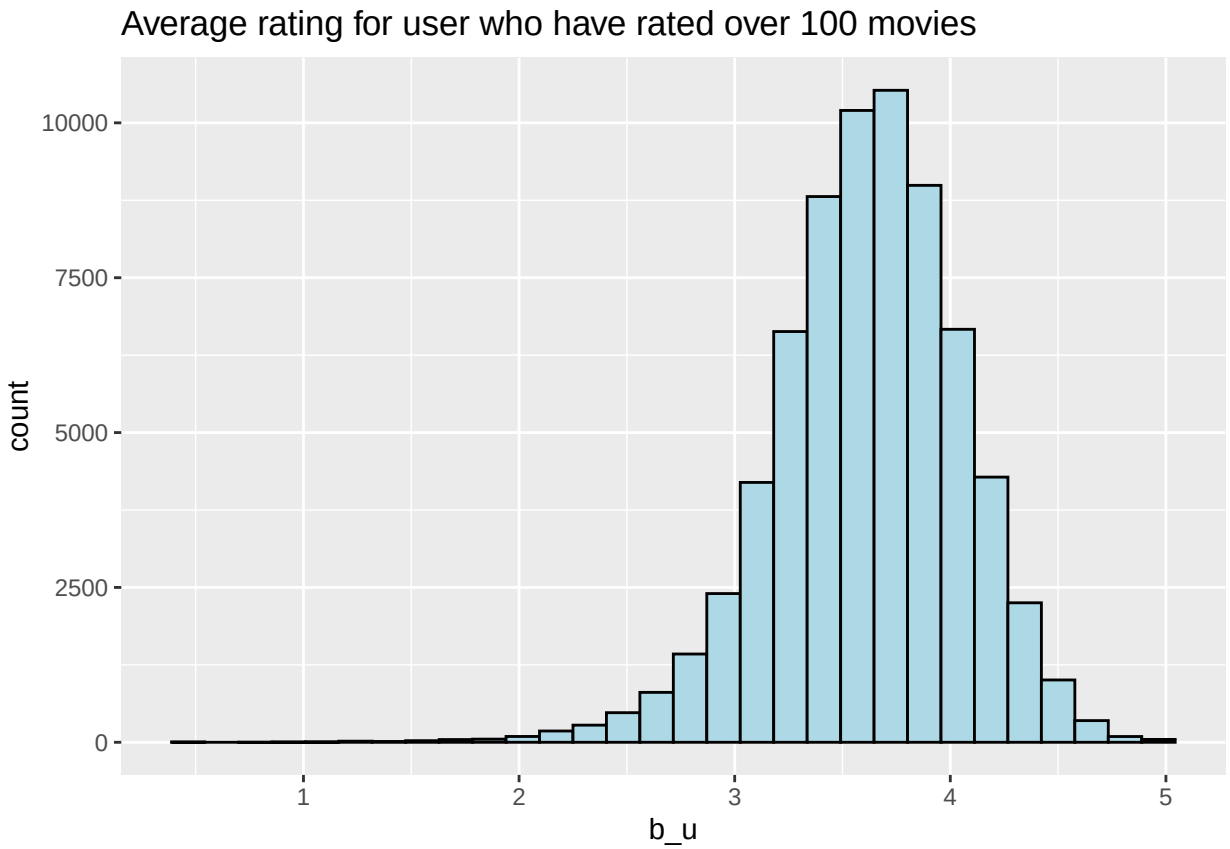
$$Y_{i,u,y,g} = \mu + b_i + \varepsilon_{i,u,y,g}$$

method	RMSE
Average rating	1.0607045
Movie Effects Model	0.9437144

The model after implemented the movie effect has a rmse 0.9437144.

2.5.2 User Effect

User effect is also one of the important bias on movie rating. Some users may give most of their viewed movies a high rating, while some are in the opposite side.



It is clearly see that some users are giving high average rating to the movie while some are giving a low average rating. A user bias term should be added to the model.

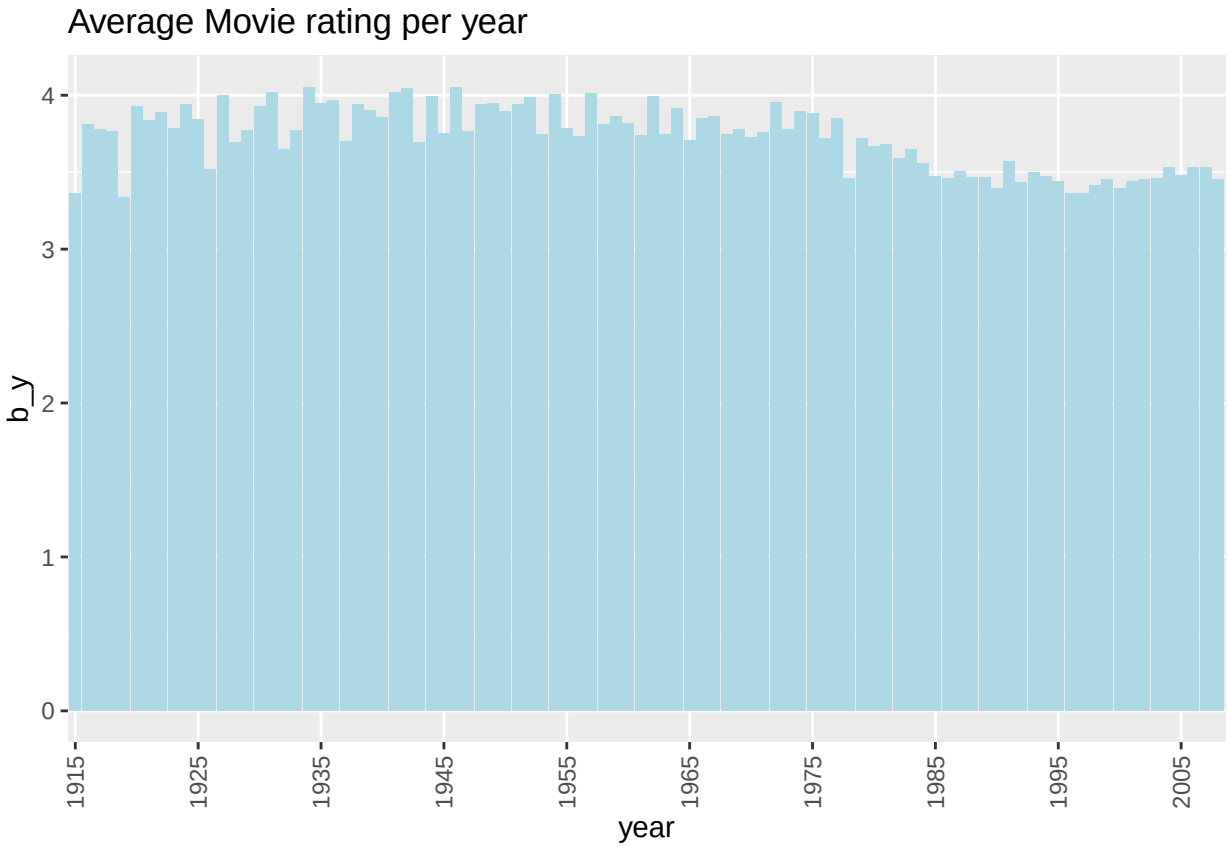
$$Y_{i,u,y,g} = \mu + b_i + b_u + \varepsilon_{i,u,y,g}$$

method	RMSE
Average rating	1.0607045
Movie Effects Model	0.9437144
Movie + User Effects Model	0.8661625

The model after implemented the user effect has improved a rmse for 0.0775519. The rmse of the current model is 0.8661625.

2.5.3 Year Effect

Movie released year could also be one of the bias. Some users may love movies that released in recent years more than old movies.



A drop of movie rating after 1975 exist. And some other years, such as 1919 and 1926, are having lower average rating. A year bias term should be added to the model.The model now become:

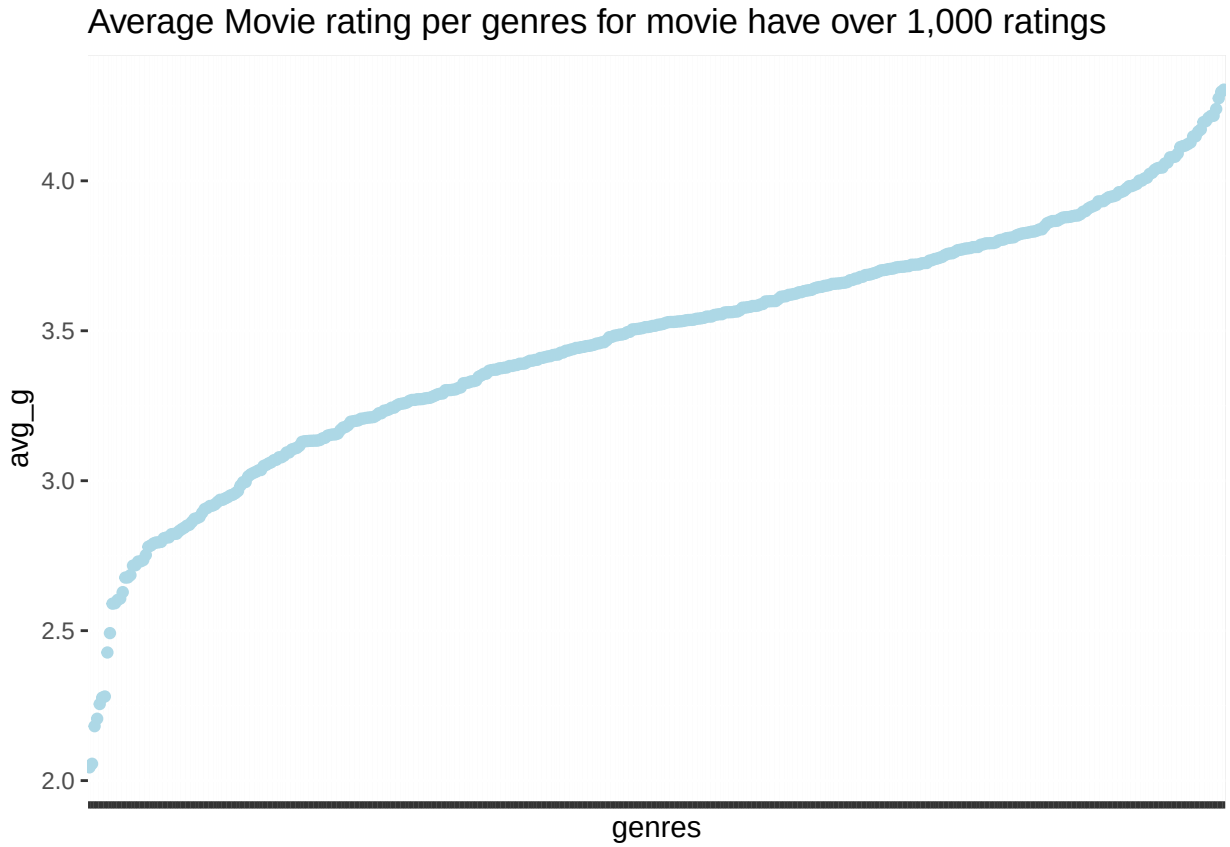
$$Y_{i,u,y,g} = \mu + b_i + b_u + b_y + \varepsilon_{i,u,y,g}$$

method	RMSE
Average rating	1.0607045
Movie Effects Model	0.9437144
Movie + User Effects Model	0.8661625
Movie + User + Year Effects Model	0.8658322

The model after implemented the year effect has improved a rmse for 0.0003303. The rmse of the current model is 0.8658322.

2.5.4 Genre Effect

Some users may love a specific genre ,such as romatic movie or comedy. Some Genre may have a higher average rating than the others.



Some genres can have an average rating over 4 while some could have only 2 or 2.5. Which the model should added a genre effect term:

$$Y_{i,u,y,g} = \mu + b_i + b_u + b_y + b_g + \varepsilon_{i,u,y,g}$$

method	RMSE
Average rating	1.0607045
Movie Effects Model	0.9437144
Movie + User Effects Model	0.8661625
Movie + User + Year Effects Model	0.8658322
Movie + User +Year + Genre Effects Model	0.8655786

The model after implemented the genre effect has improved a rmse of 0.0002536. The rmse of the current model is 0.8655786.

In generally, an exmaple of 5-star rating can now be decomposed into:

5 = 3.1 (the average rating) + 0.7 (the movie effect) + 0.9 (the user effect) + 0.7 (the year effect) - 0.4 (The genre effect)

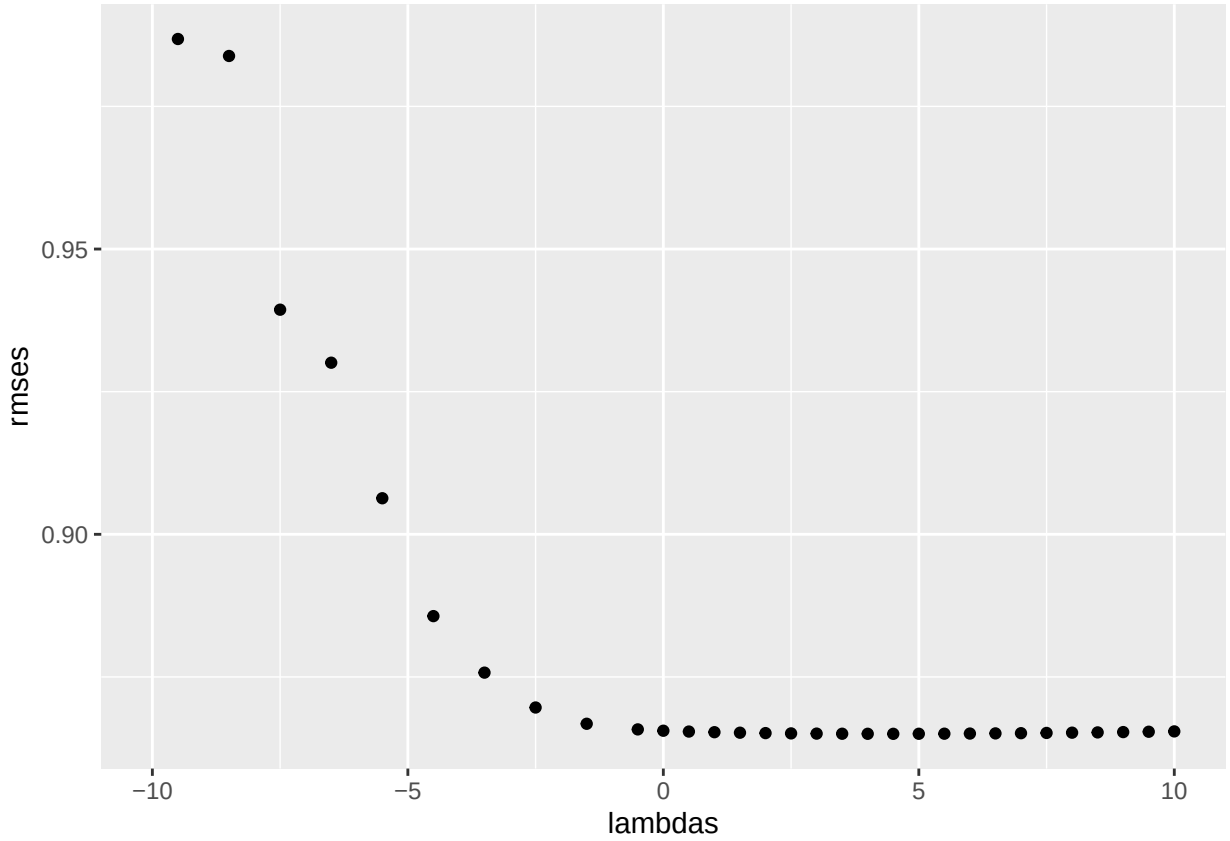
2.6 Regularization

The final rmse on the test set reported in the model is 0.8655786. To improve the accuaracy of the model, regularization approach is used:

$$\frac{1}{N}\sum_{i,u,y,g}(y_{i,u,y,g}-\mu-b_i-b_u-b_y-b_g)^2+\lambda\sum_{i,u,y,g}b_i+b_u+b_y+b_g$$

A penalty is added to the model to constrain the total variability of the effect sizes.

2.6.1 Optimal penalised coefficien




```
## [1] 4.5
```

The optimal penalised coefficient λ is found to be 4.5.

2.6.2 RMSE of the Model after regularization

method	RMSE
Average rating	1.0607045
Movie Effects Model	0.9437144
Movie + User Effects Model	0.8661625
Movie + User + Year Effects Model	0.8658322
Movie + User +Year + Genre Effects Model	0.8655786
Movie + User +Year + Genre Effects Model after regularized	0.8650367

The RMSE is improved for 0.0005419 after regularzation. The rmse become 0.8650367 after regularised.

3. Result

Model evaluation on validation set

The model has a satisfied RMSE after serveral treatments. The model is applied on the validation set for final evaluation.

method	RMSE
Movie + User +Year + Genre Effects Model after regularized	0.8642542

The final RMSE obtained from the validation set is 0.8642542.

4. Conclusion

To conclude, the model perform well in the validation set. The final rmse reported is within the acceptable range. Movie and user effects has an more important effects on the rmse of the model while released years and genres of the movie are less significant.

4.1 Limitation

Due to lack of computational power, the data set is too large too implement other machine learning algorithm ,such as KNN, random forest, directly. Other machine learning techiques, including assembled algorithm, could probably reduce the RMSE further.

4.2 Future work

For further improvement of the RMSE, martix factorization can be also applied to the model.

As movie rating prediction by residual of different effects is only part of building a movie recommendation system, a more comprehensive development on building movie recommendation system, such as implementing collaborative filtering, can be done in the future.