MovieLens_Report

Ian Gledhill

03/06/2019

```
## $value
## [1] Mean                Movie Effect Model User Effect Model
## [4] Regulated ( 5 )
## Levels: Mean Movie Effect Model User Effect Model Regulated ( 5 )
##
## $visible
## [1] FALSE
```

Introduction

Overview

This report documents the process and methodology adopted by this project to devise a movie recommendation system. The movie recommendations are derived from the predicted ratings made by a user for a given movie based on previous ratings made by the user on other films. The movie recommendtaion system will analyze a dataset of movie ratings to predict what ratings a particular user would give a movie.

The users are randomly selected and are only identified by a uniquely assigned user id; there is no demographic and geopgraphic location information used.

Dataset

Source

The source dataset was created by the University of Minnesota from the online movie recommendation service MovieLens (https://movielens.org).

It is a large dataset and contains:

- 10000054 ratings
- 95580 tags
- 10681 movies by
- 71567 users

The source files can be obtained from this location:

- https://grouplens.org/datasets/movielens/10m/

This link shows further details, including usage rights provided by the license:

- http://files.grouplens.org/datasets/movielens/ml-10m-README.html

Sample Size for Analysis

This is a very large dataset and for the sake of managability a 10% sample of the source dataset will be used by the project. The source dataset is circa 1M ratings so the sample dataset will be approximately 100K ratings.

Shape of the Data

The source data contains three data sets:

- Movies
- Ratings
- Tags

Training and Test Data Partitions

The full approach is described later in this report but in summary the sample data is split into two partitions. A *training partition* will be allocated *80%* of the sample data and will be used to create (or *train*) the algorithm. A *test partition* will be allocated the remaining *20%* of the sample data and will be used to validate the predictions of the algorithm.

Goal

The objective of the project was to find a movie recommendation algorithm. Its success will be measured by reference to the RMSE produced when the algorithm is validated against the test dataset. The RMSE will be graded against a predfined set of thresholds to determine the overall quality of the algorithm (described later).

Approach

Data Cleansing

Modelling Techniques

The algoritm will start from a naive position and then be refined as further insights into the data are gained. The effect of each refinement will be assessed using a loss function. The RMSE (*residual mean squared error*) will be used as the loss function for this project.

Version 1

Use the avarage rating of the training data to predict the rating in the test data.

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

Version 2

Use the mean with the addition of a movie effect to take into account that some movies are rated more favourably then others.

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

Version 3

Use the mean with the addition of a user effect to take into account that some users rate movies more generously then others.

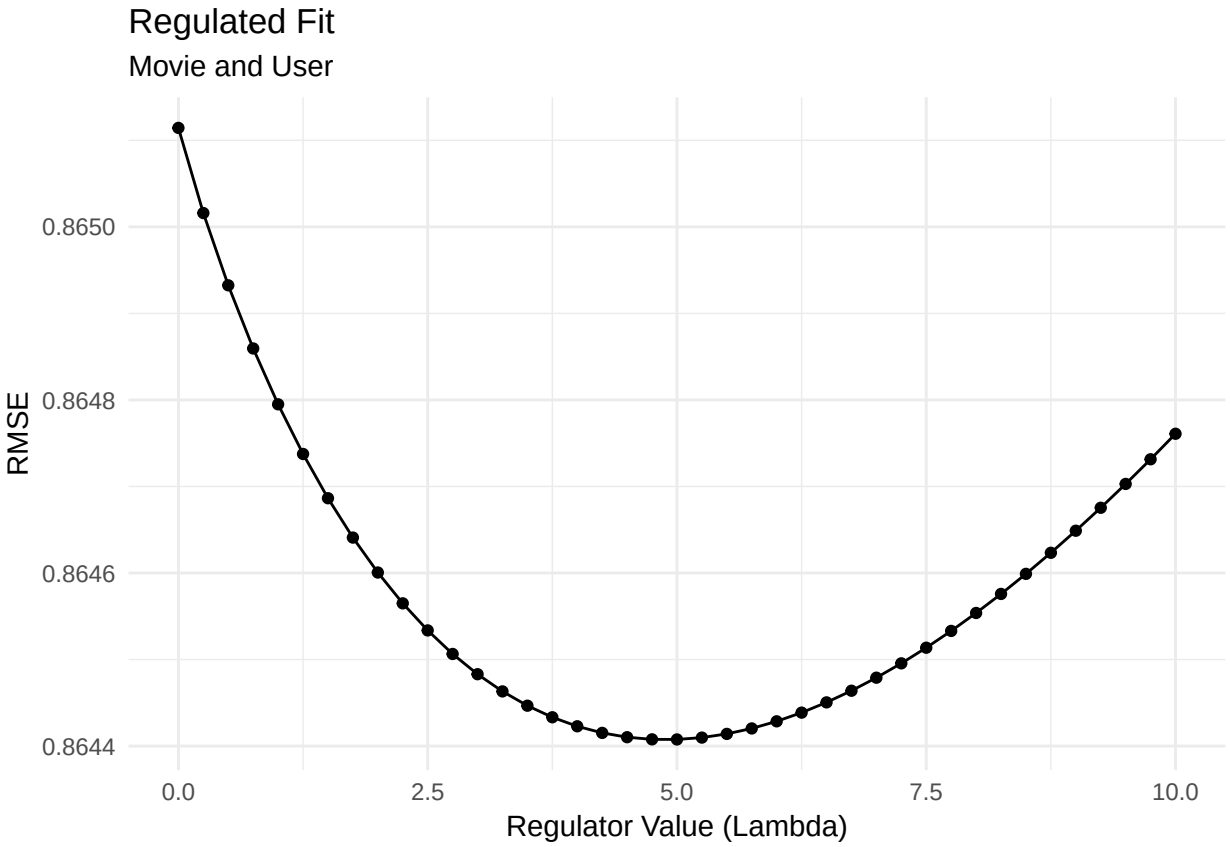
$$Y_{u,i} = \mu + b_u + \epsilon_{u,i}$$

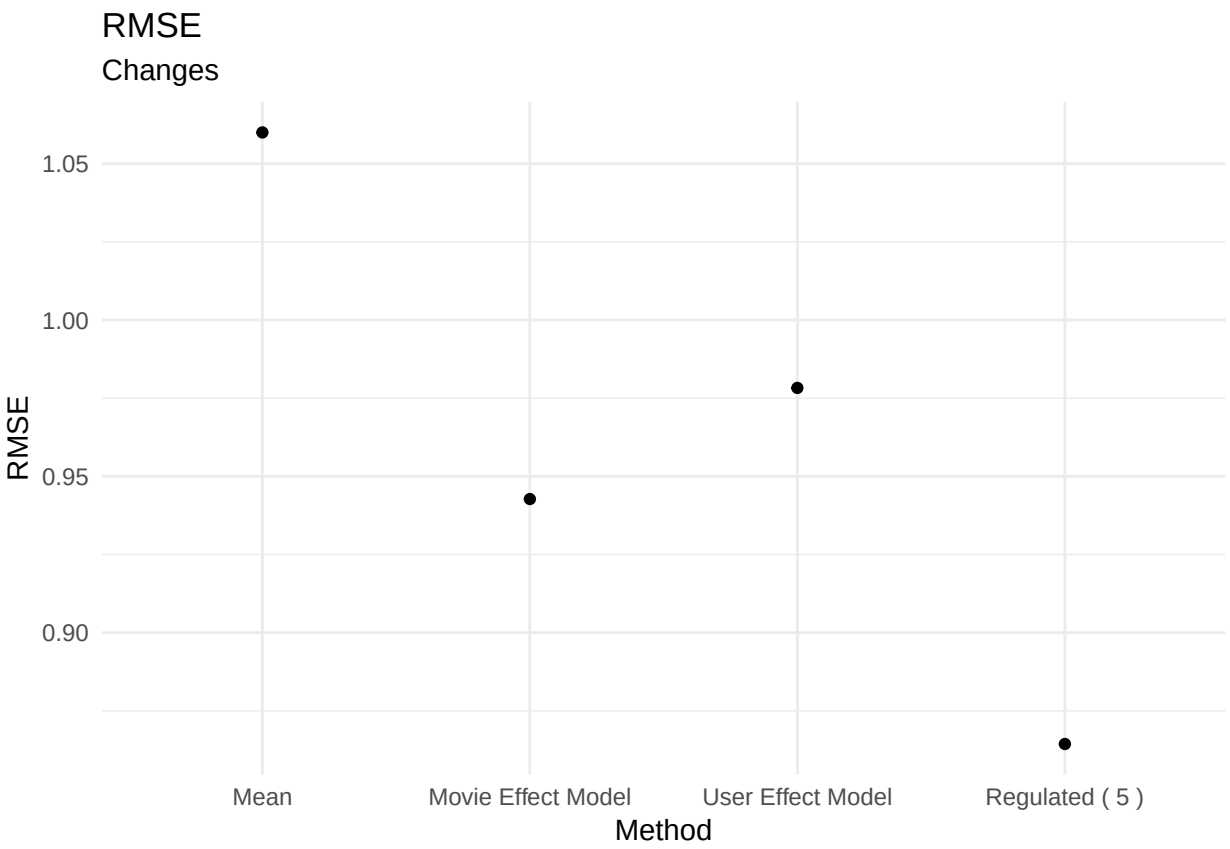
Version 4

Use the mean with the regulated movie and user effect.

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

Results


Regulated Fit
Movie and User
RMSE
0.8650
0.8648
0.8646
0.8644
0.0
2.5
5.0
7.5
10.0
Regulator Value (Lambda)


RMSE
Changes
1.05
1.00
0.95
0.90
RMSE
Mean
Movie Effect Model
User Effect Model
Regulated (5)
Method

Conclusion

Bibliography and References