MovieLens Recommender System Capstone Project

Sean Stanislaw HarvardX Professional Certificate Data Science

21/07/2021

Executive Summary

The purpose of this project is to create a movie recommendation system using the MovieLens dataset.The data set used is the 10M version of the MovieLens dataset which is divided into two parts training set and validation set. RMSE is used to test for final evaluation on the validation test. The model with the lowest Root Mean Squared Error will be selected.

Recommender systems are machine learning systems that help users discover new product and services. Every time you shop online, a recommendation system is guiding you towards the most likely product you might purchase

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

Exploratory Data Analysis

The features/variables/columns in both datasets are six:

- **userId** `<integer>` that contains the unique identification number for each user.
- **movieId** `<numeric>` that contains the unique identification number for each movie.
- **rating** `<numeric>` that contains the rating of one movie by one user. Ratings are made on a 5-Star scale with half-star increments.
- **timestamp** `<integer>` that contains the timestamp for one specific rating provided by one user.
- **title** `<character>` that contains the title of each movie including the year of the release.
- **genres** `<character>` that contains a list of pipe-separated of genre of each movie.

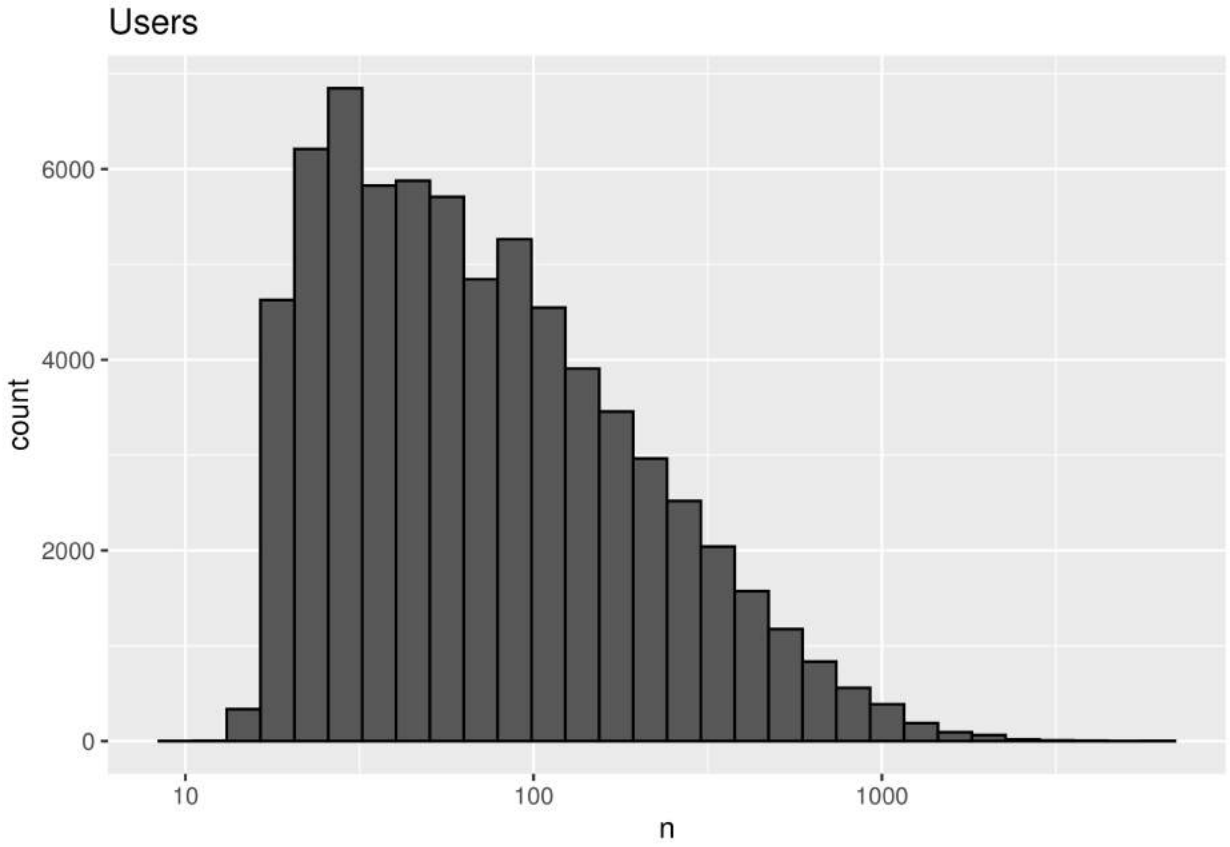
There are no missing values in the data set

```
## [1] FALSE
```

Descriptive summary of the dataset

```
##      userId         movieId          rating        timestamp
##  Min.   :    1   Min.   :    1   Min.   :0.500   Min.   :7.897e+08
##  1st Qu.:18124   1st Qu.:  648   1st Qu.:3.000   1st Qu.:9.468e+08
##  Median :35738   Median : 1834   Median :4.000   Median :1.035e+09
##  Mean   :35870   Mean   : 4122   Mean   :3.512   Mean   :1.033e+09
##  3rd Qu.:53607   3rd Qu.: 3626   3rd Qu.:4.000   3rd Qu.:1.127e+09
##  Max.   :71567   Max.   :65133   Max.   :5.000   Max.   :1.231e+09
##     title              genres
##  Length:9000055     Length:9000055
##  Class :character   Class :character
##  Mode  :character   Mode  :character
##
##
##
```

The distribution of each user's ratings for movie which demonstrates user bias



The above plot shows not every user is equally active some users rated very few movie.

First 6 Rows of edx dataset

```
##    userId movieId rating timestamp                         title
## 1:      1     122      5 838985046              Boomerang (1992)
## 2:      1     185      5 838983525               Net, The (1995)
## 3:      1     292      5 838983421               Outbreak (1995)
## 4:      1     316      5 838983392               Stargate (1994)
## 5:      1     329      5 838983392 Star Trek: Generations (1994)
## 6:      1     355      5 838984474       Flintstones, The (1994)
##                         genres
## 1:                Comedy|Romance
## 2:          Action|Crime|Thriller
## 3:  Action|Drama|Sci-Fi|Thriller
## 4:        Action|Adventure|Sci-Fi
## 5: Action|Adventure|Drama|Sci-Fi
## 6:        Children|Comedy|Fantasy
```

Total movie ratings per genre

```
## # A tibble: 6 x 2
##   genres                count
##   <chr>                 <int>
## 1 Drama                733296
## 2 Comedy               700889
## 3 Comedy|Romance       365468
## 4 Comedy|Drama         323637
## 5 Comedy|Drama|Romance 261425
## 6 Drama|Romance        259355
```

Analysis - Model Building and Evaluation

The Simple Model

The formula used is:

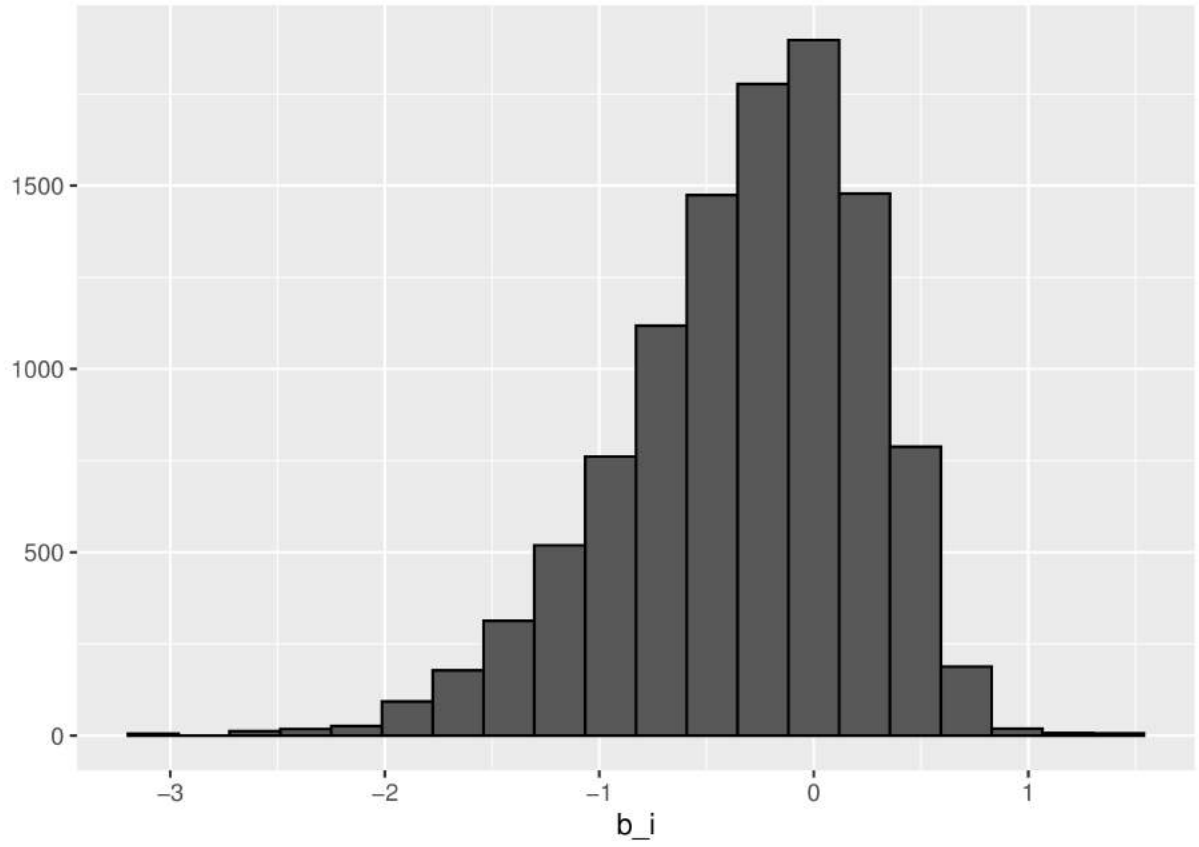
$$Y_{u,i} = \hat{\mu} + \varepsilon_{u,i}$$

With $\hat{\mu}$ is the mean and $\varepsilon_{i,u}$ is the independent errors sampled from the same distribution centered at 0.

```
## [1] 3.512465
```

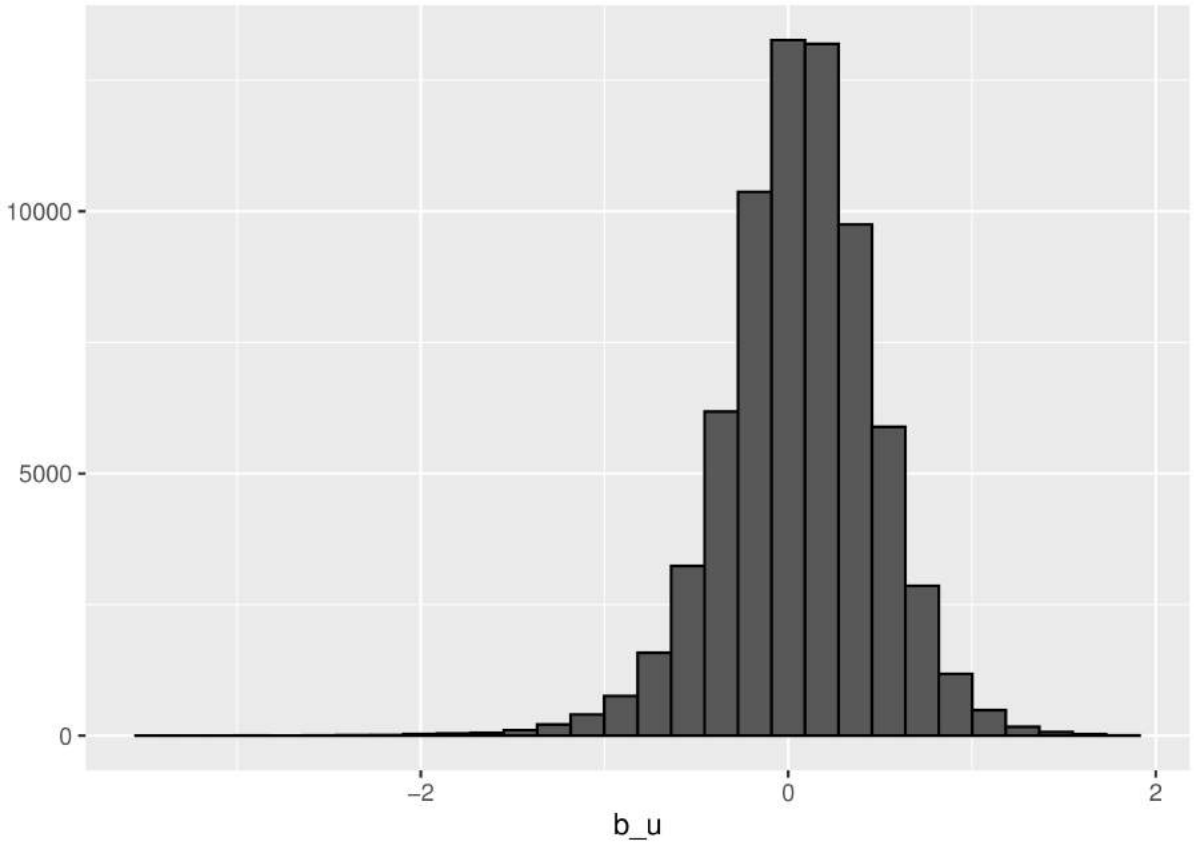
Penalty Term (b_i)- Movie Effect

Different movies have different rating as shown the histogram is not symmetric and is skewed towards negative rating effect. The movie effect is the difference from mean rating.



Penalty Term (b_u)- User Effect

Every User rates different movie differently. Some may give poor rating to a good movie and vice- versa the plot below demosntarte teh User effect



Baseline Mode

This model simply calculates mean rating . The model acts as a baseline on which we will try to improve RMSE relative to standard model

```
## [1] 1.061202
```

```
## Warning: `data_frame()` was deprecated in tibble 1.1.0.
## Please use `tibble()` instead.
```

Movie Effect Model

The RMSE is improved by adding movie effect.

method	RMSE
Using mean only	1.0612018
Movie Effect Model	0.9439087

Movie and User Effect Model

Considering movie and users biases both affect the prediction RMSE is further improved by adding user effect

method	RMSE
Using mean only	1.0612018
Movie Effect Model	0.9439087
Movie and User Effect Model	0.8653488

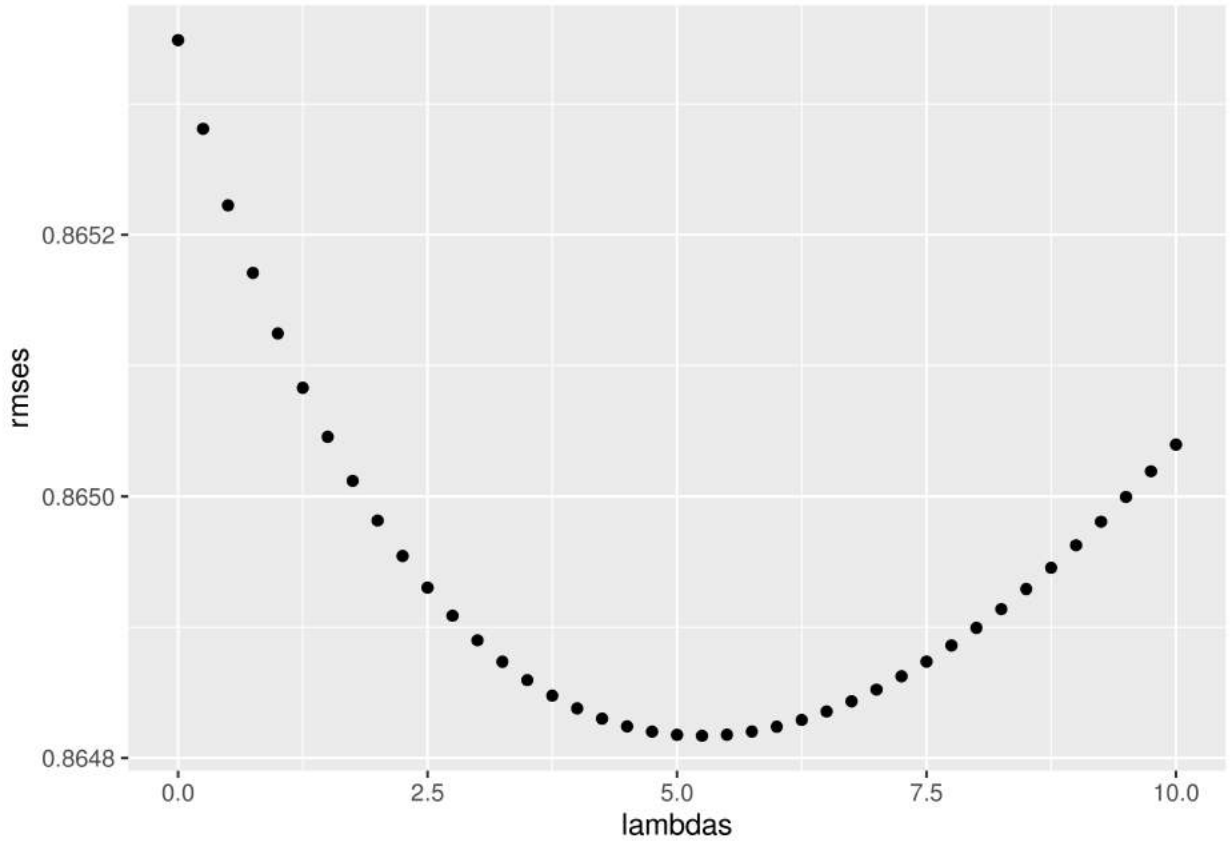
Regularization based approach.

RMSE are sensitive to large errors. Large errors can increase our residual mean squared error. So we must put a penalty term to give less importance to such effect. e regularization method allows us to add a penalty λ (lambda) to penalizes movies with large estimates from a small sample size. In order to optimize b_i, it necessary to use this equation:

$$\frac{1}{N}\sum_{u,i}(y_{u,i}-\mu-b_i)^2+\lambda\sum_i b_i^2$$

reduced to this equation:

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




```
## [1] 5.25
```

The minimum Lambda value is 5.25

Calculate RMSE for Regularization based approach.

method	RMSE
Regularized Movie and User Effect Model	0.864817

Concluding Remarks

The RMSE table shows Using mean only RMSE is 1.061 , only movie effect is RMSE is 1.0612 , Movie Effect Model RMSE is 0.9439 Movie and User Effect Mode RMSE is 0.8653 and the final model Regularized Movie and User Effect Model RMSE is 0.8648. So the best permoring model is teh regualrisation model which provides the lowest RMSE and the model that will be selected .

Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.