JS_MovieLens

Ioannis Skiadas

2023-12-01

Contents

INTRODUCTION

In 2009 Netflix announced the winners of its competition for an improved recommendation software (A Feuerverger 2012). The metric used in the competition was the reduction of the Residual Mean Square Error (RMSE). However Netflix did not make the used data set public. Instead, GroupLens Labs generated their own data set comprising 27000 movies and 138000 users (Irizarry 2020). A subset of this data set, MovieLens wll be used in the present work to develop a recommendation software able to produce a RMSE less than 0.86490. The system will be built gradually taking into account how the movies where rated as well as the effect of different raters and if needed the effects of additional features such as the movie Genres. Moreover, in order to account for the different number of rating among different movies a penalized approach will also be used. The project's documentation will be implemented through R markdown (Xie 2023).

METHODS/ANALYSIS

Software

The software used for the analysis will be R R version 4.1.0 (2021-05-18) through RStudio 2022.07.2+576.

Dataset and packages

The dataset used as well as the necessary packages will be available through the following:

The function for the estimation of the predicted rating error loss, will be the Residual Mean Square Error (RMSE):

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

where *y'* is the predicted rating of user u for movie i whereas y is the observed rating.

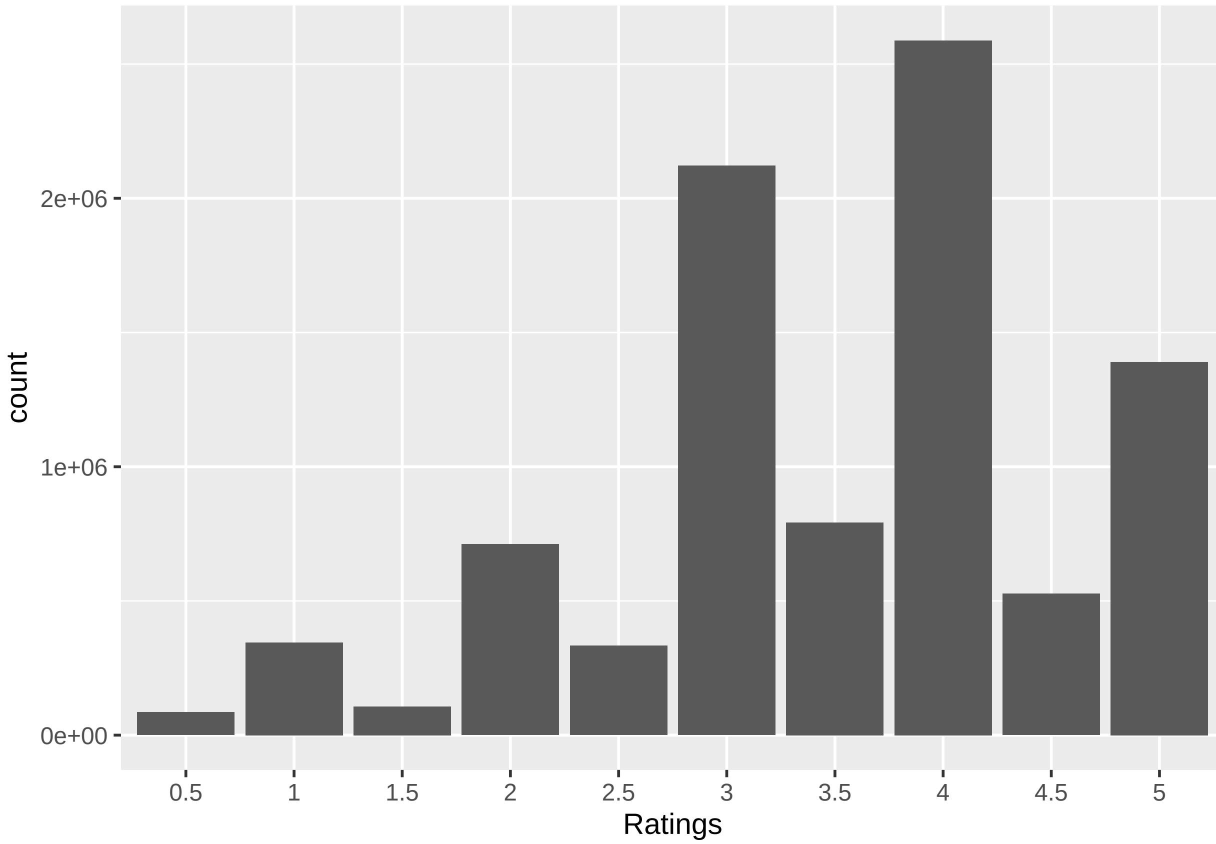
Overall, a summary of the Dataset is:

```
##      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
##
##
##
```

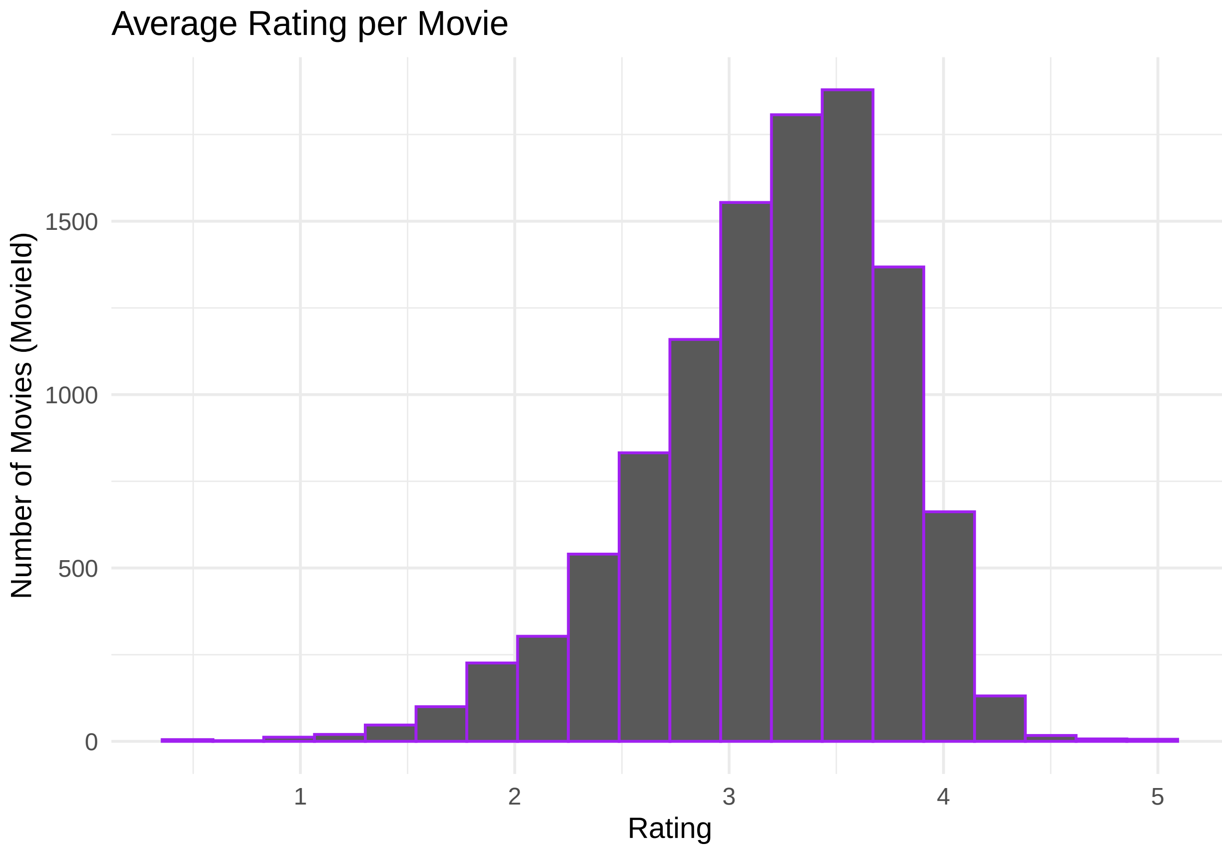
whereas the set comprises:

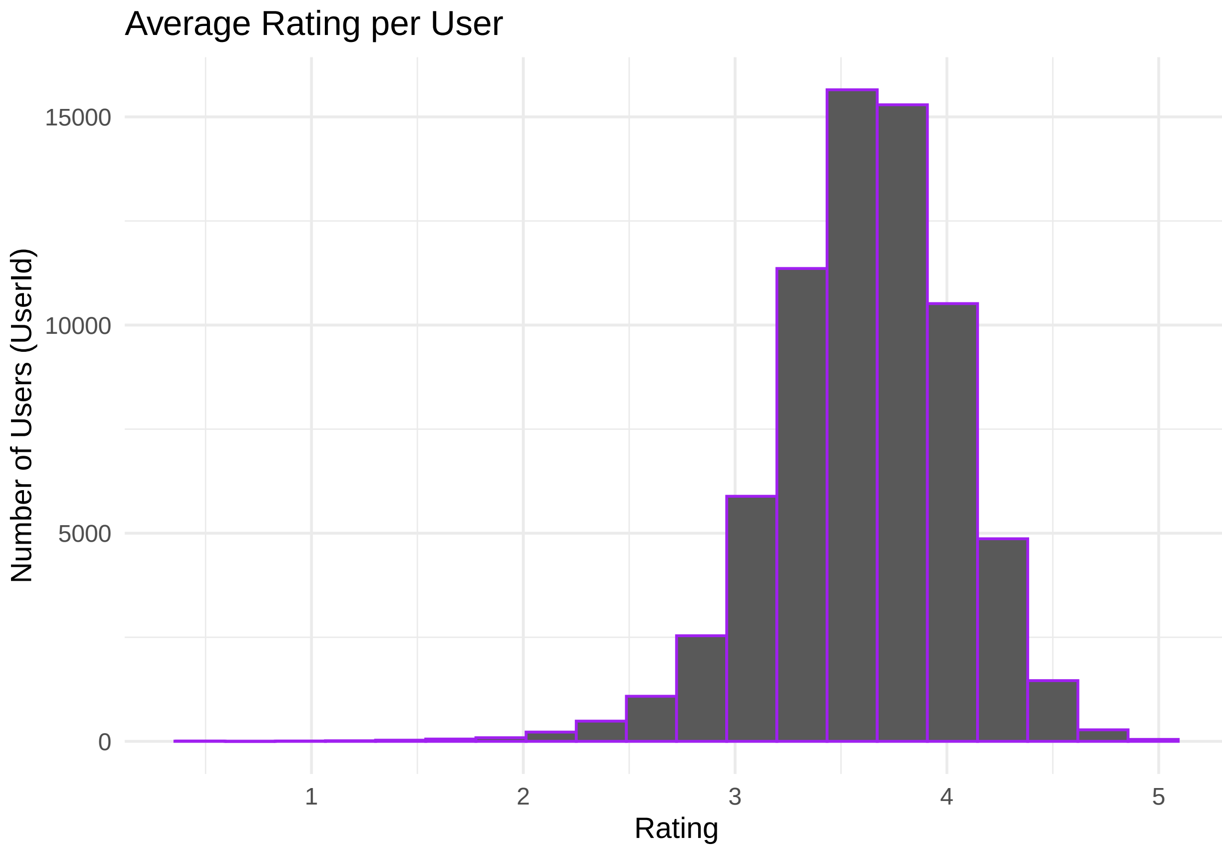
Users	Movies	Genres
69878	10677	797

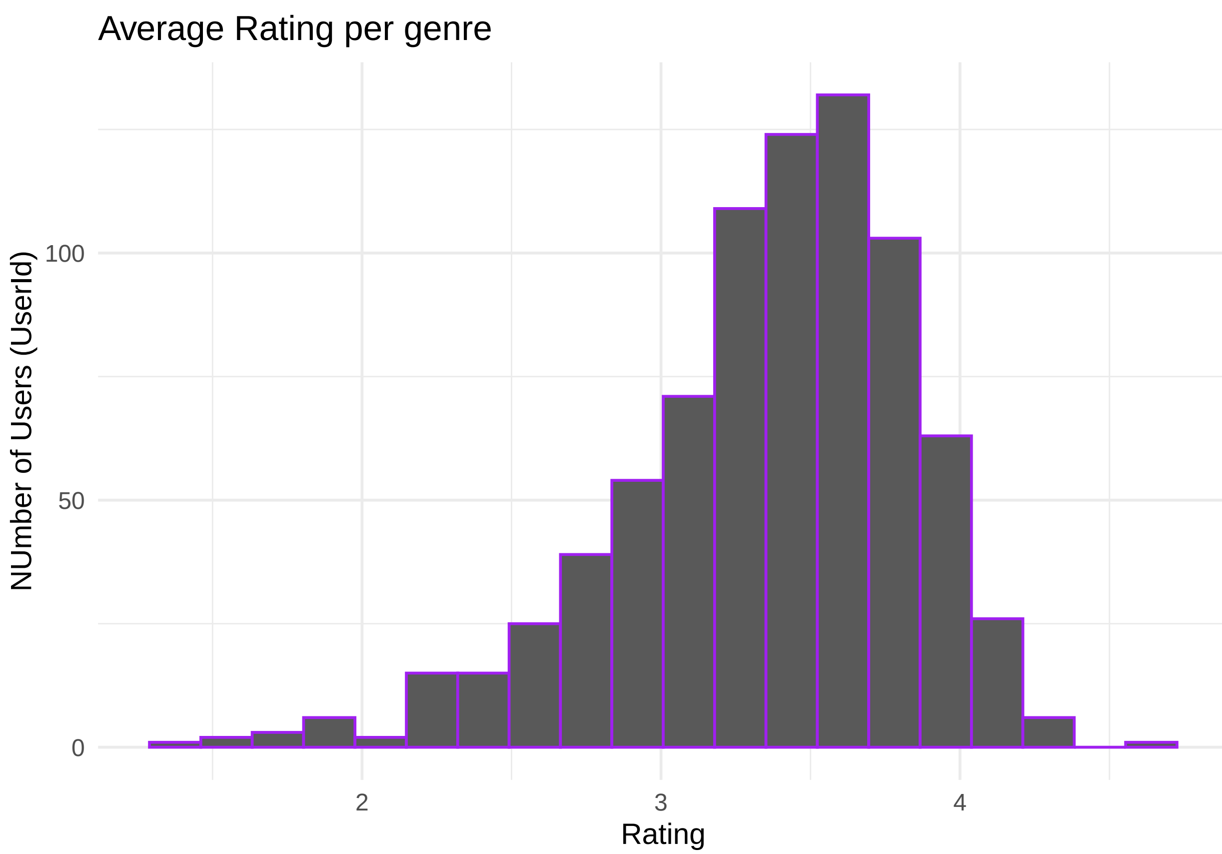
The rating Grades are distributed as:



A further Description of the distributions of the average rating for movies, users and gendre categories are:


Average Rating per Movie
Number of Movies (MovieId)
1500
1000
500
0
1
2
3
4
5
Rating


Average Rating per User
Number of Users (UserId)
15000
10000
5000
0
1
2
3
4
5
Rating



Whereas, the top rated movies are:

Table 2: Top Rated Movies

movieId	title	mean
3226	Hellhounds on My Trail (1999)	5.00
33264	Satan's Tango (Sátántangó) (1994)	5.00
42783	Shadows of Forgotten Ancestors (1964)	5.00
51209	Fighting Elegy (Kenka erejii) (1966)	5.00
53355	Sun Alley (Sonnenallee) (1999)	5.00
64275	Blue Light, The (Das Blaue Licht) (1932)	5.00
5194	Who's Singin' Over There? (a.k.a. Who Sings Over There) (Ko to tamo peva) (1980)	4.75
26048	Human Condition II, The (Ningen no joken II) (1959)	4.75
26073	Human Condition III, The (Ningen no joken III) (1961)	4.75
65001	Constantine's Sword (2007)	4.75

Moreover, since the MovieLens dataset involves a large number of records and parameters - thousands or records and movies - whose combinations would render regression calculations very long i.e. a *linear regression model, lm*, such an approach will not be followed. Instead, an estimation of the model coefficients will be followed as the average of the difference of the predicted rating from the mean rating of each movie, had the effects not been taken into account i.e: for the movie effect least square estimate bm:

$$Y_{u,i} - \mu'$$

the user coefficient once their effect is taken into account as well, bu:

$$Y_{u,i} - bm - \mu'$$

or with the addition of the Genre effect: or

$$Y_{u,i} - bm - bu - \mu'$$

,

and the regularization factor will be estimated, i.e. for the movie, user and Genres effects:

$$\frac{1}{N}\sum_{u,i}(y_{u,i} - \mu - b_m - b_u - b_g)^2 + \lambda(\sum_m b_m^2 + \sum_u b_u^2 + \sum_g b_g^2)$$

Finally, it is important to note that concluding the training steps, the model will be evaluated against the *final_holdout_test* which it had not seen before for the final outcome.

RESULTS

Preparation of the training and test dat sets.

Assessing the average rating of the training set.

The RMSE function which will be used for the evaluation of the different models:

A first approach would be to suggest the average of the ratings as an approximation. Its error loss is estimated below

Table 3: RESULTS

Model	RMSE
Naive	1.061135

The same approach could be also taken for the median. Its own error presented below:

Table 4: RESULTS

Model	RMSE
Naive	1.061135
Naive_Median	1.167939

As the error terms does not improve, the effect of the features are estimated. Accounting for the movie effect:

Table 5: RESULTS

Model	RMSE
Naive	1.0611350
Naive_Median	1.1679394
Movie	0.9441568

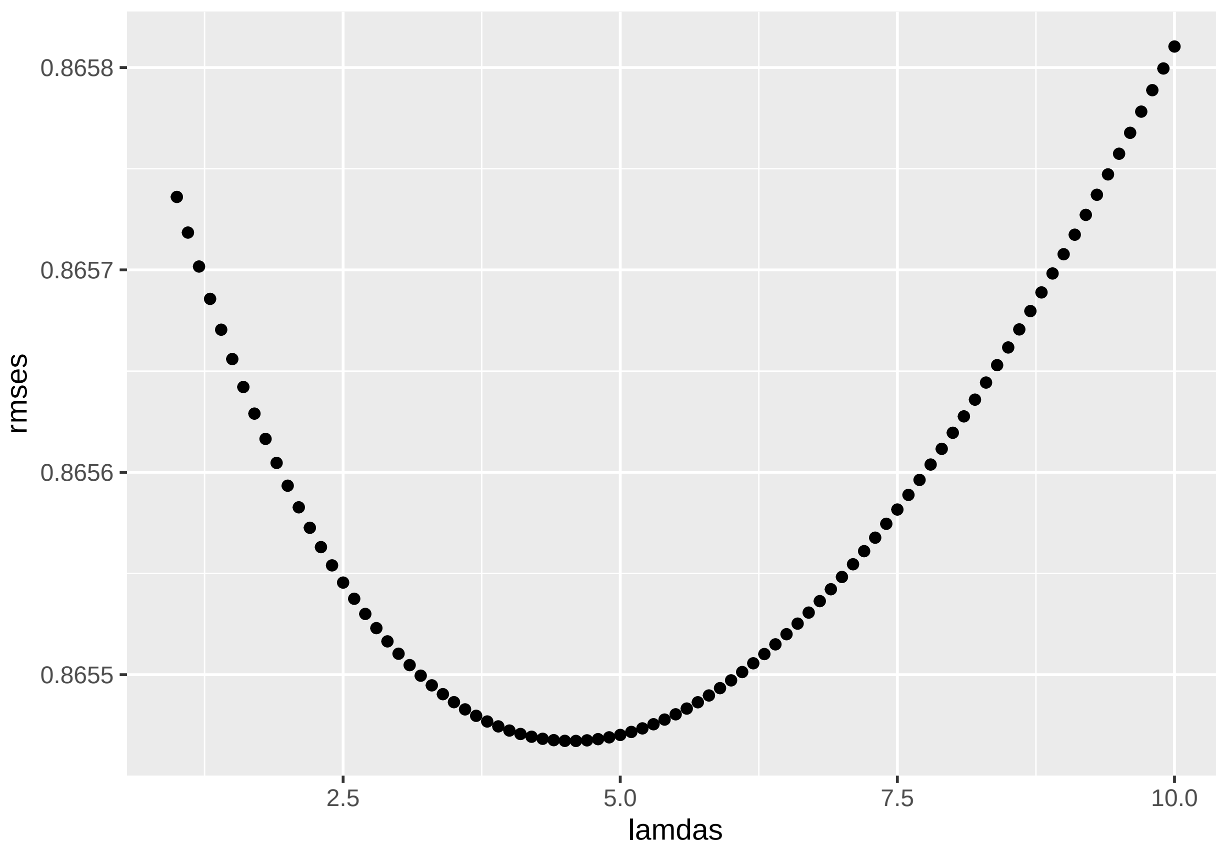
Adding the user effect:

Table 6: RESULTS

Model	RMSE
Naive	1.0611350
Naive_Median	1.1679394
Movie	0.9441568
Movie and User	0.8659736

Estimating the penalty term, through cross validation, to evaluate the training of the model so far

The distribution of the estimated lamdas:



And the value with the lowest RMSE

```
## [1] 4.6
```

Adding it to the developing model:

Table 7: RESULTS

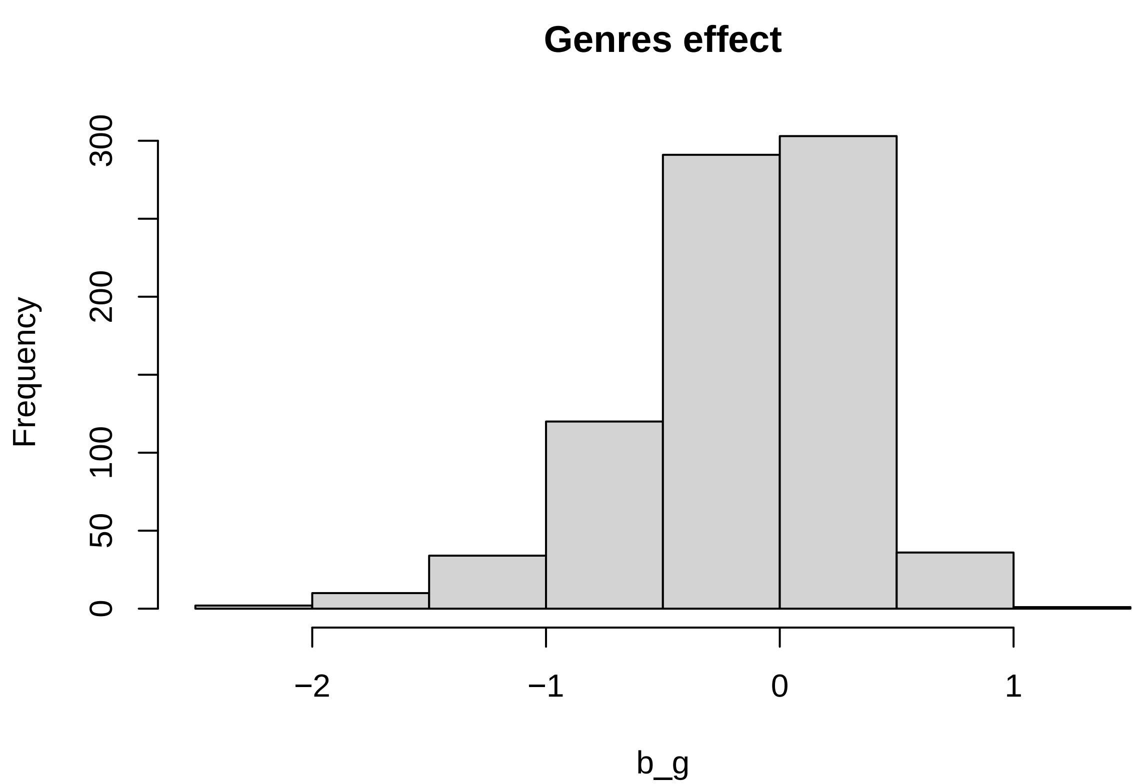
Model	RMSE
Naive	1.0611350
Naive_Median	1.1679394
Movie	0.9441568
Movie and User	0.8659736
RMSE_regularized	0.8654673

The resulting prediction

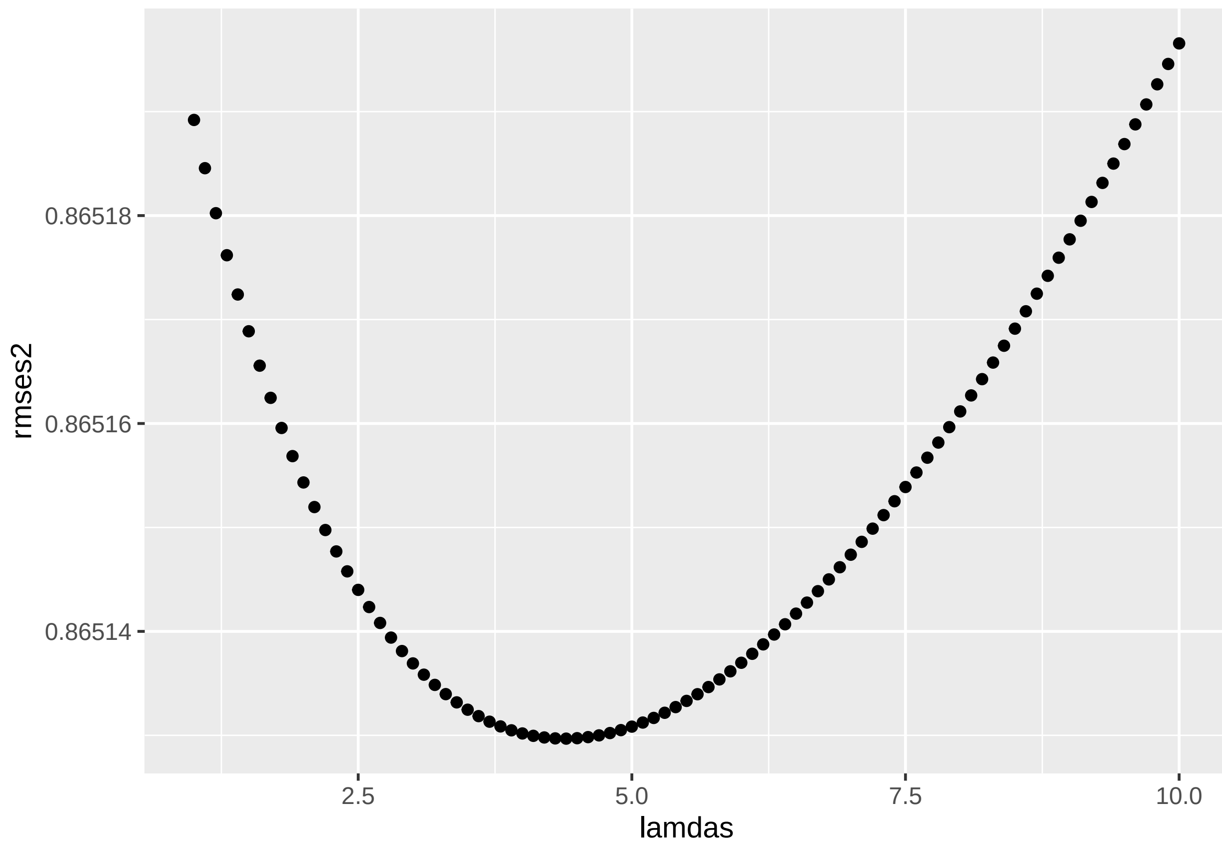
Table 8: RESULTS

Model	RMSE
Naive	1.0611350
Naive_Median	1.1679394
Movie	0.9441568
Movie and User	0.8659736
RMSE_regularized	0.8654673
RMSE_regularized_First_Test	0.8652065

Since this is not sufficient (less than 0.8649), the effect of the Genres is also taken into account



The new Lamda's distribution



And the updated value:

```
## [1] 4.6
```

Resulting in an improved model:

Table 9: RESULTS

Model	RMSE
Naive	1.0611350
Naive_Median	1.1679394
Movie	0.9441568
Movie and User	0.8659736
RMSE_regularized	0.8654673
RMSE_regularized_First_Test	0.8652065
RMSE_M_U_G_Regularized	0.8651298

And its final overall evaluation againgt an unseen data set

Table 10: RESULTS

Model	RMSE
Naive	1.0611350
Naive_Median	1.1679394
Movie	0.9441568
Movie and User	0.8659736
RMSE_regularized	0.8654673
RMSE_regularized_First_Test	0.8652065
RMSE_M_U_G_Regularized	0.8651298
RMSE_FInal_regularized	0.8648367

CONCLUSION

The added effect of the different Genre Categories was sufficient to reduce the error loss at the desired levels,*<0.8649*, specifically at 0.8648367.

REFERENCES

A Feuerverger, S Khatri, Y He. 2012. "Statistical Significance of the Netflix Challenge." *Statistical Science* 27, 202-231.

Irizarry, Rafael A. 2020. *Introduction to Data Science.* HarvardX Data Science Series. https://rafalab.github.io/dsbook/.

Xie, Yihui. 2023. *Bookdown:authoring Books and Technical Documents with r Markdown.* CRC Press. https://bookdown.org/yihui/bookdown/.