MovieLens document

Robert van der Heijden

2/23/2020

Content

1. Summary section: describes the dataset and summarizes the goal of the project and key steps that were performed
2. Analysis section: explains the process and techniques used, including data cleaning, data exploration and visualization, insights gained, and your modeling approach
3. Results section: presents the modeling results and discusses the model performance
4. Conclusion section: gives a brief summary of the report, its limitations and future work

1) Summary section

For this project, we will be creating a movie recommendation system using the MovieLens dataset. The full version of movielens includes millions of ratings. We will use the 10M version of the MovieLens dataset to make the computation a little easier.

The goal of this project is to train a machine learning algorithm using the inputs in the 10M version of the MovieLens dataset. The predictions from this machine learning algorithm will be compared to the true ratings in the validation set using RMSE.

2) Analysis Section

The first step is to create an EDX set and validation set.

Create edx set, validation set

Data Exploration (Before Data Cleaning)

What is in the Edx data set?

1. The first ten lines of the dataset:

userId	movieId	rating	timestamp	title	genres
1	122	5	838985046	Boomerang (1992)	Comedy|Romance
1	185	5	838983525	Net, The (1995)	Action|Crime|Thriller
1	231	5	838983392	Dumb & Dumber (1994)	Comedy
1	292	5	838983421	Outbreak (1995)	Action|Drama|Sci-Fi|Thriller
1	316	5	838983392	Stargate (1994)	Action|Adventure|Sci-Fi
1	329	5	838983392	Star Trek: Generations (1994)	Action|Adventure|Drama|Sci-Fi
1	355	5	838984474	Flintstones, The (1994)	Children|Comedy|Fantasy
1	356	5	838983653	Forrest Gump (1994)	Comedy|Drama|Romance|War
1	362	5	838984885	Jungle Book, The (1994)	Adventure|Children|Romance
1	364	5	838983707	Lion King, The (1994)	Adventure|Animation|Children|Drama|Musical

2. The summary of the statistics of the dataset:

	userId	movieId	rating t	imestamp	title	genres
Min	. : 1 Min	. : 1 Min	. :0.500 Min	. :7.897e+08 Len	gth:9000061 Len	gth:9000061
1st	Qu.:18122 1st	Qu.: 648 1st	Qu.:3.000 1st	Qu.:9.468e+08 Cla	ss :character Cla	ss :character
Med	ian :35743 Med	ian : 1834 Med	ian :4.000 Med	ian :1.035e+09 Mod	e :character Mod	e :character
Mea	n :35869 Mea	n : 4120 Mea	n :3.512 Mea	n :1.033e+09 NA	NA	
3rd	Qu.:53602 3rd	Qu.: 3624 3rd	Qu.:4.000 3rd	Qu.:1.127e+09 NA	NA	
Max	. :71567 Max	. :65133 Max	. :5.000 Max	. :1.231e+09 NA	NA	

3. The number of rows is:	9000061
4. The number of columns is:	6
5. How many different movies are in the edx dataset?	10677
6. How many different users are in the edx dataset?	69878

7. Which movie has the greatest number of ratings?

title	number
Pulp Fiction (1994)	31336
Forrest Gump (1994)	31076
Silence of the Lambs, The (1991)	30280
Jurassic Park (1993)	29291
Shawshank Redemption, The (1994)	27988
Braveheart (1995)	26258
Terminator 2: Judgment Day (1991)	26115
Fugitive, The (1993)	26050
Star Wars: Episode IV - A New Hope (a.k.a. Star Wars) (1977)	25809
Batman (1989)	24343

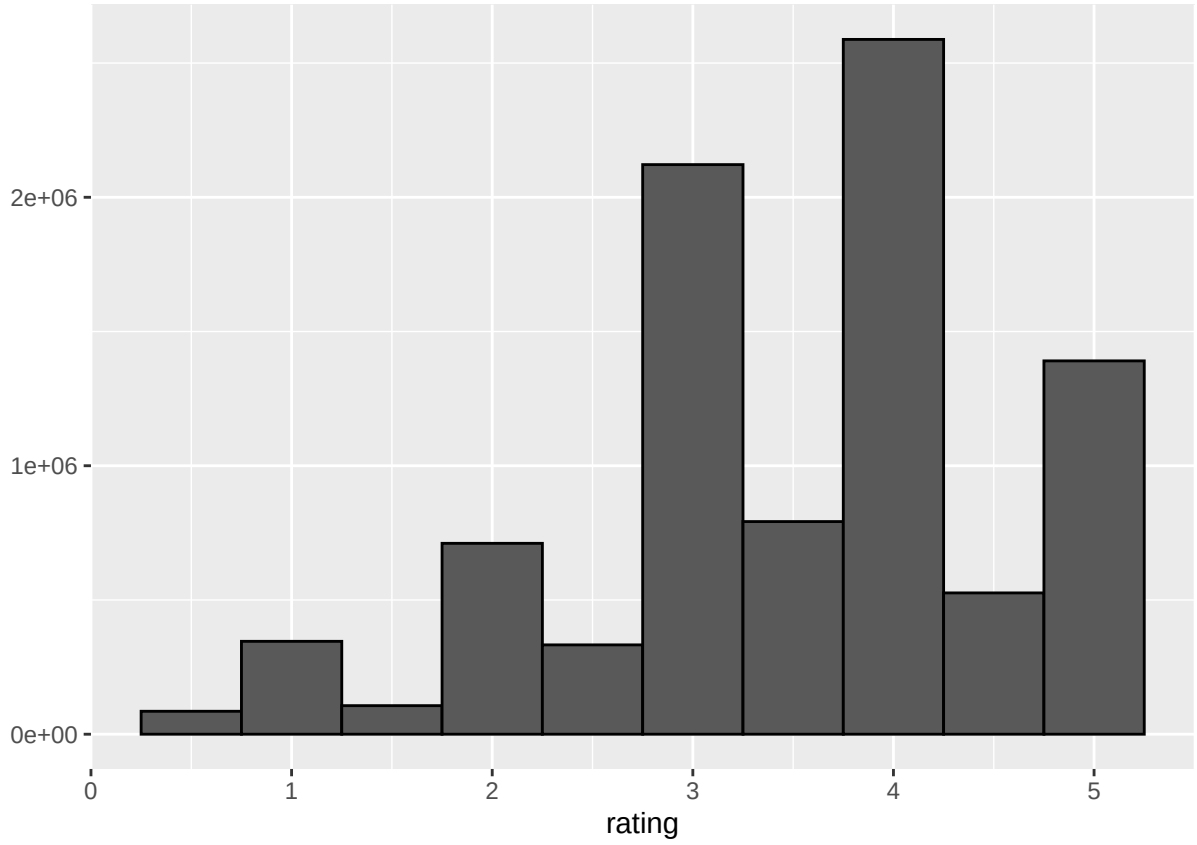
Data Cleaning

- Modify the year as a column in the edx & validation datasets

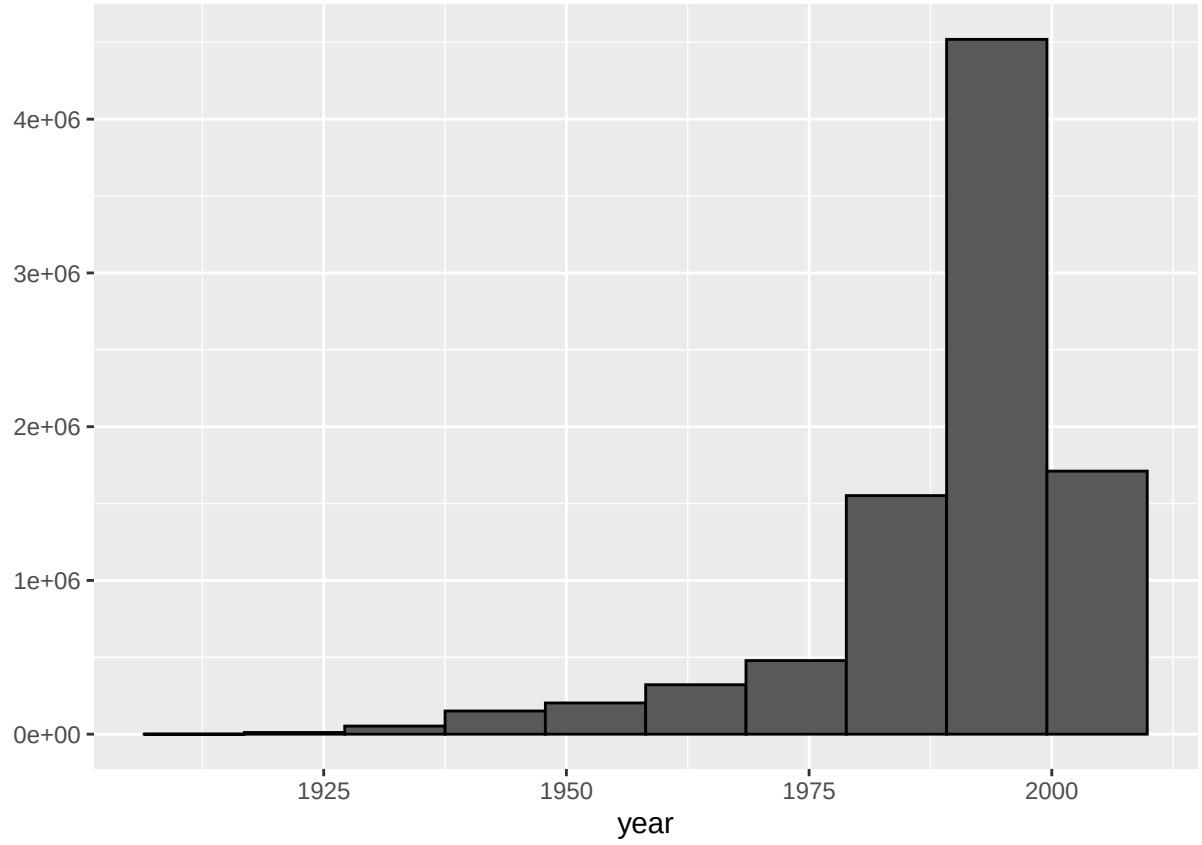
Year added to edx set, validation set

Data Exploration (After Data Cleaning)

How are the ratings distributed (histogram)?



How are the ratings distributed over the years (histogram)?



Modelling Approach

- RMSE will be used to evaluate how close the predictions are to the true values in the validation set.

Building the recommendation system

1. We start by building the simplest possible recommendation system. We're going to predict the same rating for all movies. The average that we predict is:

3.512464

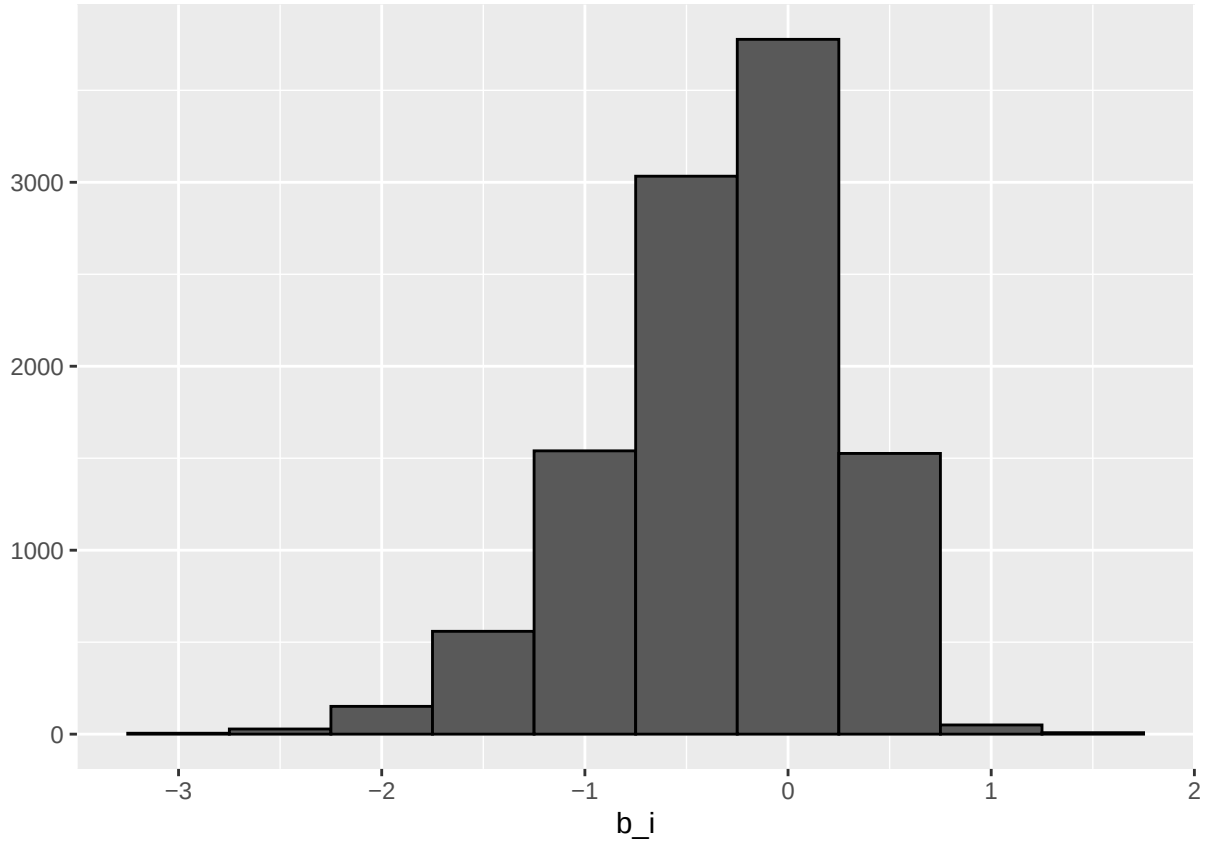
How well does this model? The deviation on average is:

1.0606506

Now because as we go along we will be comparing different approaches, we're going to create a table that's going to store the results that we obtain as we go along.

Method	RMSE
Just the average	1.060651

2. In the second step, we are going to take the movie effect into account.

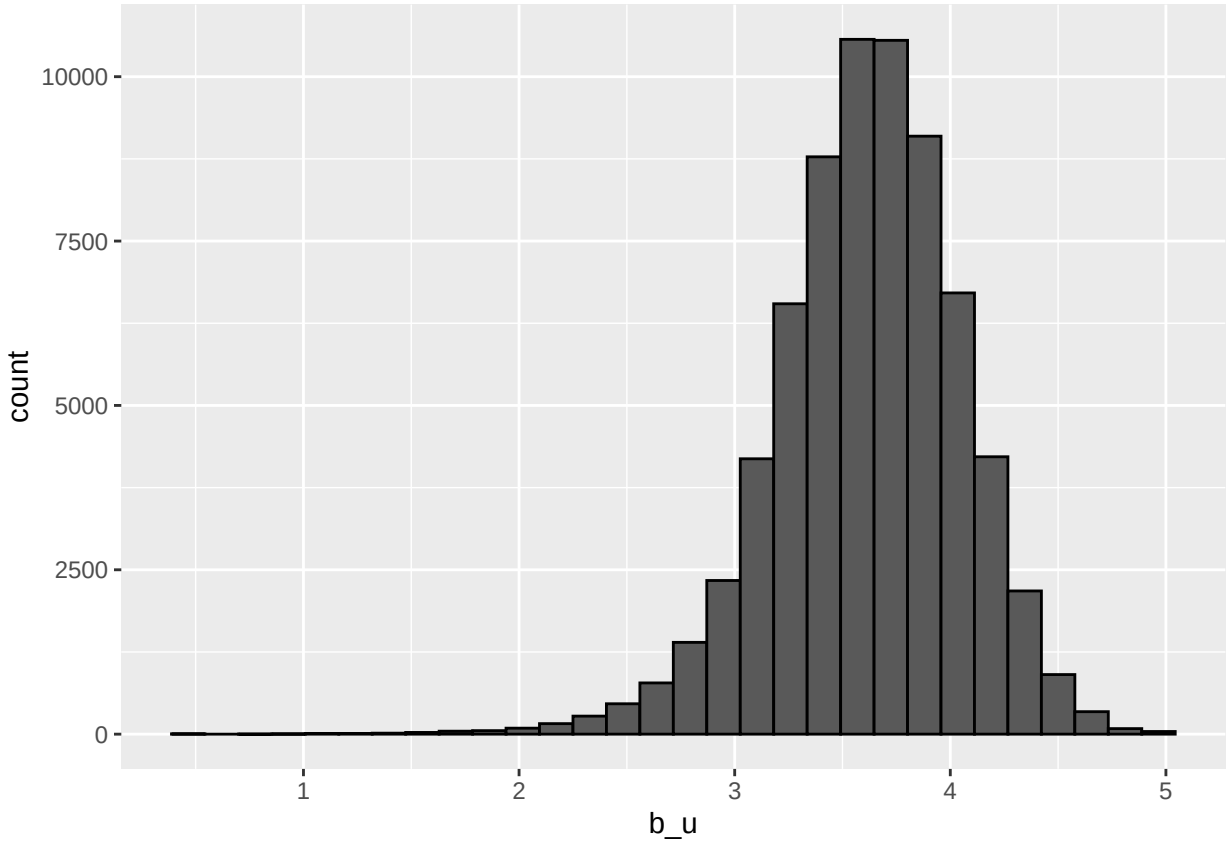


The histogram shows the deviation of a movie rating from the average rating. You can see that these estimates vary substantially. Some movies are good, other movies are bad.

Updated RMSE table:

Method	RMSE
Just the average	1.0606506
Movie Effect Model	0.9437046

3. In the third step, we are going to take the movie effect and user effect into account.



The histogram shows the variability of a movie rating of a user. There is substantial variability across users

Updated RMSE table:

Method	RMSE
Just the average	1.0606506
Movie Effect Model	0.9437046
Movie + User Effects Model	0.8655329

4. in the fourth step, we are going to regularize the movie and user effect.

Best and worst movies are rated by a few users, but have a big impact on our model. We use regularization to improve our model

Top 5 Best predicted movies:

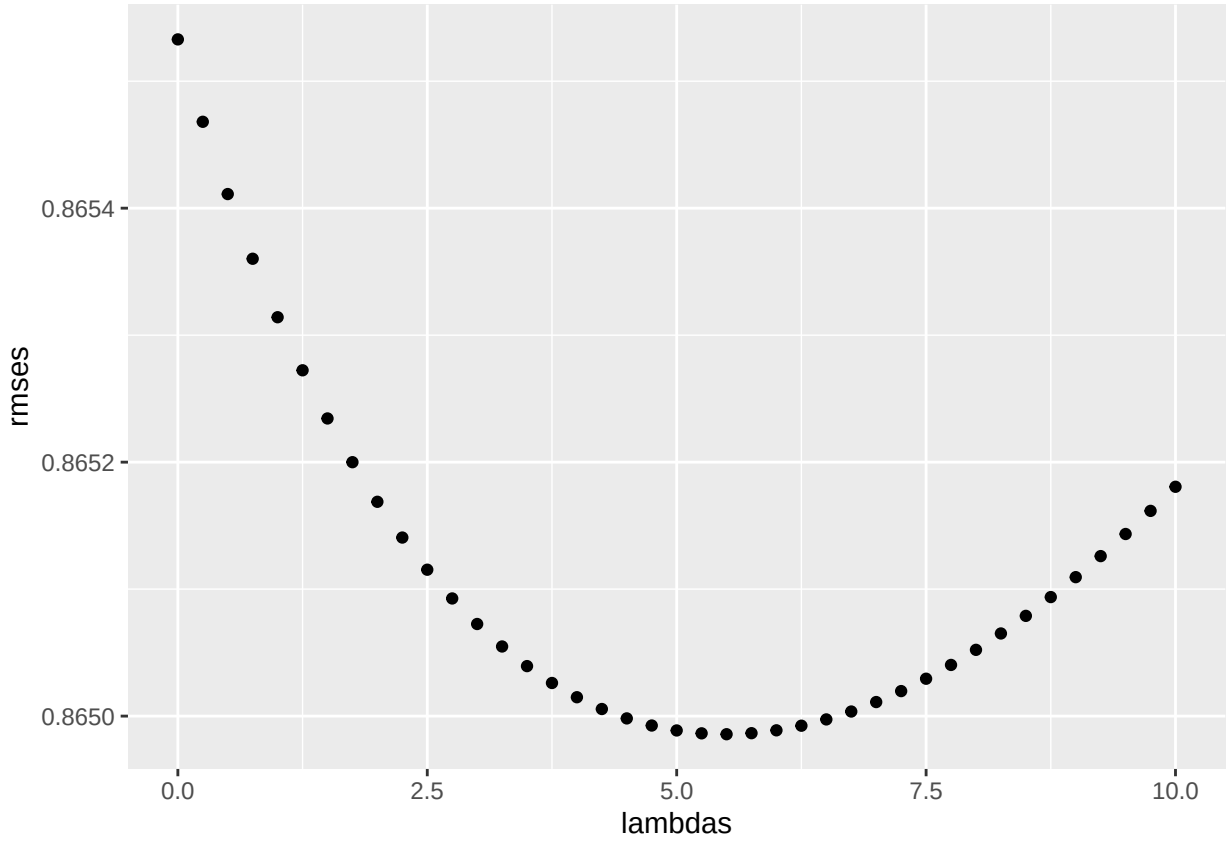
title	b_i	n
Hellhounds on My Trail (1999)	1.487536	1
Satan's Tango (SÃ¡tÃ¡ntangÃ³) (1994)	1.487536	2
Shadows of Forgotten Ancestors (1964)	1.487536	1
Fighting Elegy (Kenka erejii) (1966)	1.487536	1
Sun Alley (Sonnenallee) (1999)	1.487536	1

Top 5 Worst predicted movies:

title	b_i	n
Besotted (2001)	-3.012464	2
Hi-Line, The (1999)	-3.012464	1
Grief (1993)	-3.012464	1
Accused (Anklaget) (2005)	-3.012464	1
War of the Worlds 2: The Next Wave (2008)	-2.762464	2

Large errors can increase our residual mean squared error, so we would rather be conservative when we're not sure. Regularization permits us to penalize large estimates that come from small sample sizes.

First we have to pick the optimal tuning parameter lambda.



The optimal lambda is:

```
## [1] 5.5
```

Updated RMSE table:

Method	RMSE
Just the average	1.0606506
Movie Effect Model	0.9437046
Movie + User Effects Model	0.8655329
Regularized Movie + User Effects Model	0.8649857

3) Results Section

RMSE overview

The RMSE values for the used models are shown below:

Method	RMSE
Just the average	1.0606506
Movie Effect Model	0.9437046
Movie + User Effects Model	0.8655329
Regularized Movie + User Effects Model	0.8649857

The RMSE table shows an improvement of the model over the different assumptions. The simplest model 'Using mean only' calculates a RMSE of more than 1.06, which means, on average, we miss the rating by one star. Incorporating 'Movie effect' and 'Movie and User effect' in our model gives an improvement of 11% and 18.4% respectively.

A deeper insight into the data revealed some data points have large effect on errors. So a regularization model was used to penalize these kind of data points. The final RMSE is 0.864986 with an improvement over 18.45% with respect to the baseline model.

Other sources of variation (e.g. the fact that groups of movies have similar rating patterns and groups of users have similar rating patterns as well) can be added to the model to further improve the predictability of the model.

4) Conclusion Section

For this project, we created a movie recommendation system using the MovieLens dataset. The full version of movielens includes millions of ratings. We used the 10M version of the MovieLens dataset to make the computation a little easier.

The goal of this project is to train a machine learning algorithm using the inputs in the 10M version of the MovieLens dataset. The predictions from this machine learning algorithm were compared to the true ratings in the validation set using RMSE.

We started with a simple model, using the 'mean rating' only, and added 'Movie Effects' and 'User Effects'. Our final model included regularization, that improved the final RMSE to 0.864986. This was an improvement over 18.45% with respect to the baseline model.

Other sources of variation (e.g. the fact that groups of movies have similar rating patterns and groups of users have similar rating patterns as well) can be added to the model to further improve the predictability of the model.