MovieLens Project

Gonzalo Vazquez

02-23-2023

Executive Summary

The dataset used in this project is MovieLens 10M dataset, which is a dataset of movie ratings collected by MovieLens, a movie recommendation service. The goal of this project is to predict movie ratings from edx dataset using the final_holdout_test dataset. The key steps performed in this project are loading and pre-processing the dataset, making data exploration, modeling and showing some conclusions from there.

Exploring the dataset

First lets to explore some of the rows and columns the dataset has

	userId	movieId	rating	timestamp	title	genres
1	1	122	5	838985046	Boomerang (1992)	Comedy|Romance
2	1	185	5	838983525	Net, The (1995)	Action|Crime|Thriller
4	1	292	5	838983421	Outbreak (1995)	Action|Drama|Sci-Fi|Thriller
5	1	316	5	838983392	Stargate (1994)	Action|Adventure|Sci-Fi
6	1	329	5	838983392	Star Trek: Generations (1994)	Action|Adventure|Drama|Sci-Fi
7	1	355	5	838984474	Flintstones, The (1994)	Children|Comedy|Fantasy

Then we can check if there are some missing values in the dataset to see how to proceed for the analysis

Table 2: Missing Values by Variable

	Values
userId	0
movieId	0
rating	0
timestamp	0
title	0
genres	0

We can also check the quantity of users, movies, ratings and genres

Table 3: Quantity of unique values of each variable

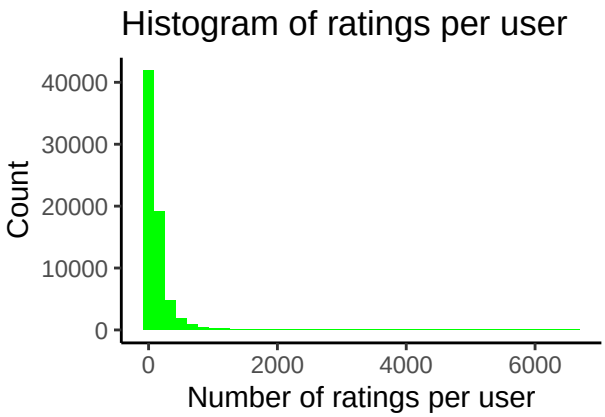
Users	Movies	Ratings	Genres
69878	10677	10	797

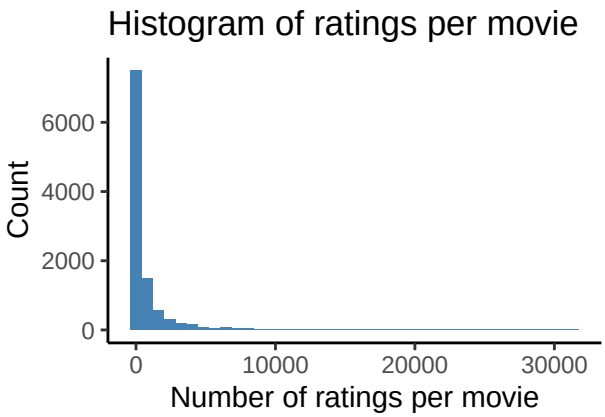
Here we can see the different genres and how many there are

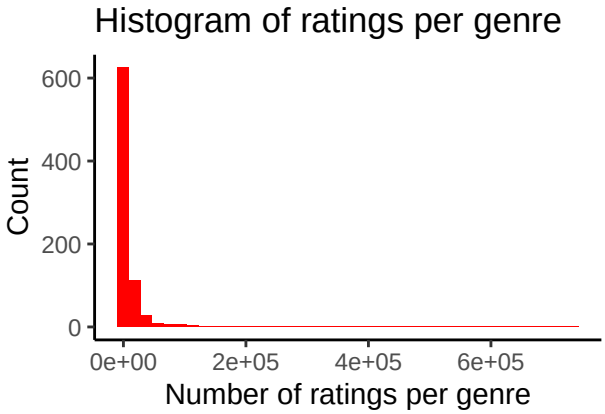
Table 4: Genres Summary

Genre	Count
Comedy	3540930
Romance	1712100
Action	2560545
Crime	1327715
Thriller	2325899
Drama	3910127
Sci-Fi	1341183
Adventure	1908892
Children	737994
Fantasy	925637
War	511147
Animation	467168
Musical	433080
Western	189394
Mystery	568332
Film-Noir	118541
Horror	691485
Documentary	93066
IMAX	8181
(no genres listed)	7

Finally I will check the number of ratings across different dimensions of the dataset: users, movies and genres







We can see that there are a lot of ratings concentrated in a few movies and genres. We can also see that there are users that rated movies more than others

Analysis

To predict the ratings of an user for a movie we will use this formula:

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

Where $Y_{u,i}$ represent the predicted rating of user u for item i. The terms μ, b_i, and b_u represent, respectively, the overall mean rating, the bias associated with item i, and the bias associated with user u. Finally, $\epsilon_{u,i}$ represents the error term for the predicted rating $Y_{u,i}$.

To evaluate the accuracy of recommendation models we will use this function:

```
calculate_rmse <- function(real_ratings, ratings_predictions){
  sqrt(mean((real_ratings - ratings_predictions)^2))
}
```

We can start calculating the RMSE between the actual ratings (edx$rating) and the mean rating (mu_hat). Then we will keep adding more dimensions to see if we can reduce the RMSE result

Table 5: RMSE Summary

Method	RMSE
average	1.060331

Movie effect model incorporates biases associated with movies. We can observe a lower RMSE with the movie effect

Table 6: RMSE Summary

Method	RMSE
Movie Effect Model	0.9439087

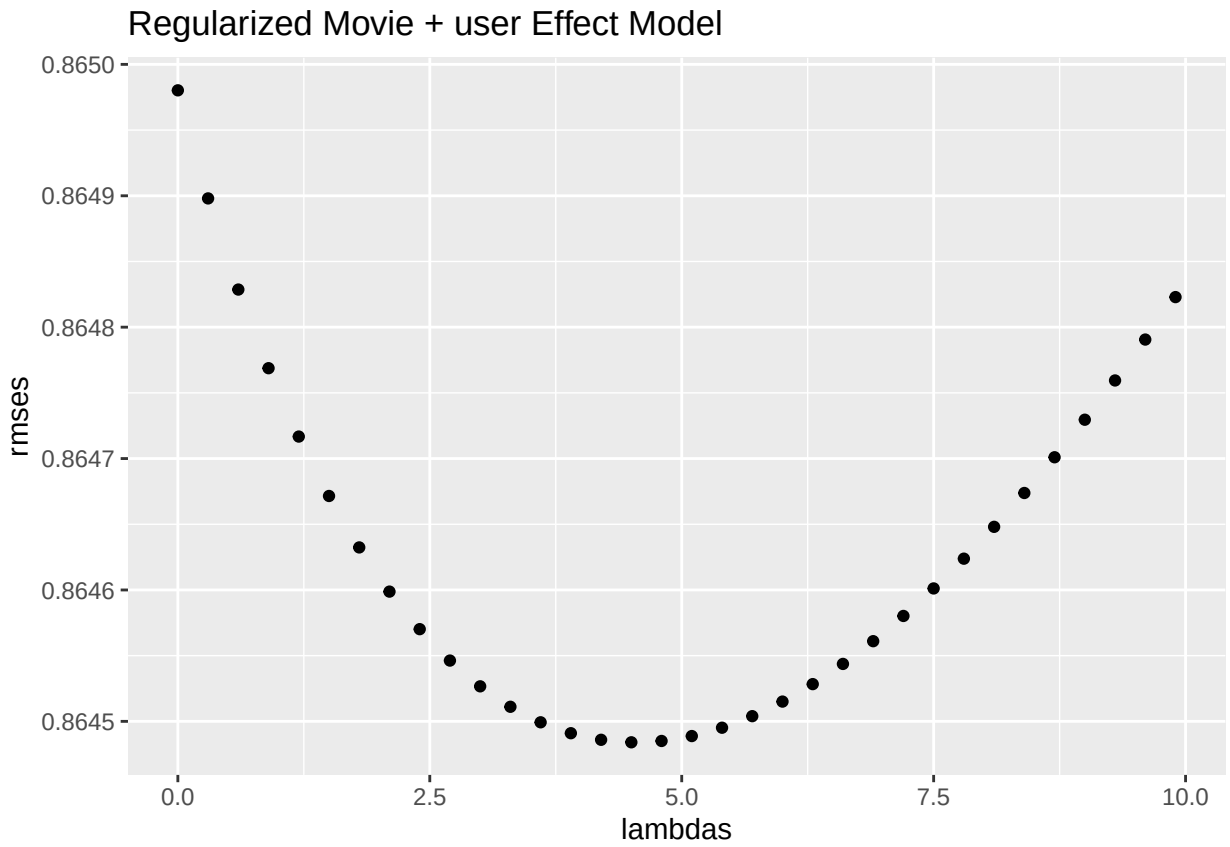
User effect model is built on the previous model, adding the biases associated with users. We can observe that the RMSE is also reduced applying this model

Method	RMSE
Movie + User Effects Model	0.8653488

Here movie, user, and genre effect model is developed, which includes biases associated with genres in addition to the previous two. The RMSE result is also reduced with this model

Method	RMSE
Movie + User + Genre Effect Model	0.8647516

Here is the regularized movie and user effect model. It includes an additional parameter λ that shrinks the biases to the mean of the biases.



Method	RMSE
Regularized movie + user effect model	0.8648242

Final results:

As is shown in the data exploration histograms, we can observe that a few movies and genres receive a large number of ratings, while a few people tend to rate many movies. Therefore, in the analysis section, I worked with different models to evaluate how the prediction results improve, as I show in the table below::

Method	RMSE
Average	1.0603
Movie Effect Model	0.9439
Movie + User Effect Model	0.8653
Movie + User + Genre Effect Model	0.8647
Regularized movie + user effect model	0.8648

Conclusions:

We can conclude that working with models with more dimensions increase the precision of the results but we can't conclude that regularized Movie+User Effect model improves the final results as it has a slightly higher RMSE compared with the Movie+User+Genre Effect Model.

Overall, the results suggest that the Movie + User + Genre Effect Model is the best-performing method among the ones compared, with an RMSE of 0.8647

For further research we could explore why regularization don't decrease the RMSE and keep experimenting with different regularization strengths and evaluate their RMSE