MovieLens Project - Report

Corentin Lobet

1. Introduction

Goal of this research

This work is the first project of the final examination required to achieve the HarvardX Data Science Program. Analyses are performed on the MovieLens dataset as the goal is to compute a well performing recommendation movie system. The machine learning models used here are linear models predicting ratings based on the measurement of biases. We measure the performance of our models with the RMSE.

The MovieLens Data

We used a subpart of the MovieLens data containing 10 millions observations. This data can be downloaded here :

- MovieLens 10M dataset
- MovieLens 10M dataset - zip file

We will work on 90% of this data for most of the analyses in order to keep away 10% of the data as a validation set for our models.

2. Analysis

Description of the data variables

In our analysis we use the ratings given by users as the outcome to predict. We use three variables as predictors :

- movieId : an ID number unique to each movie
- userId : an ID number given to each user
- genres : genre associated to the movie

The data also include the titles and the date and time (as timestamp) the rating was given.

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	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

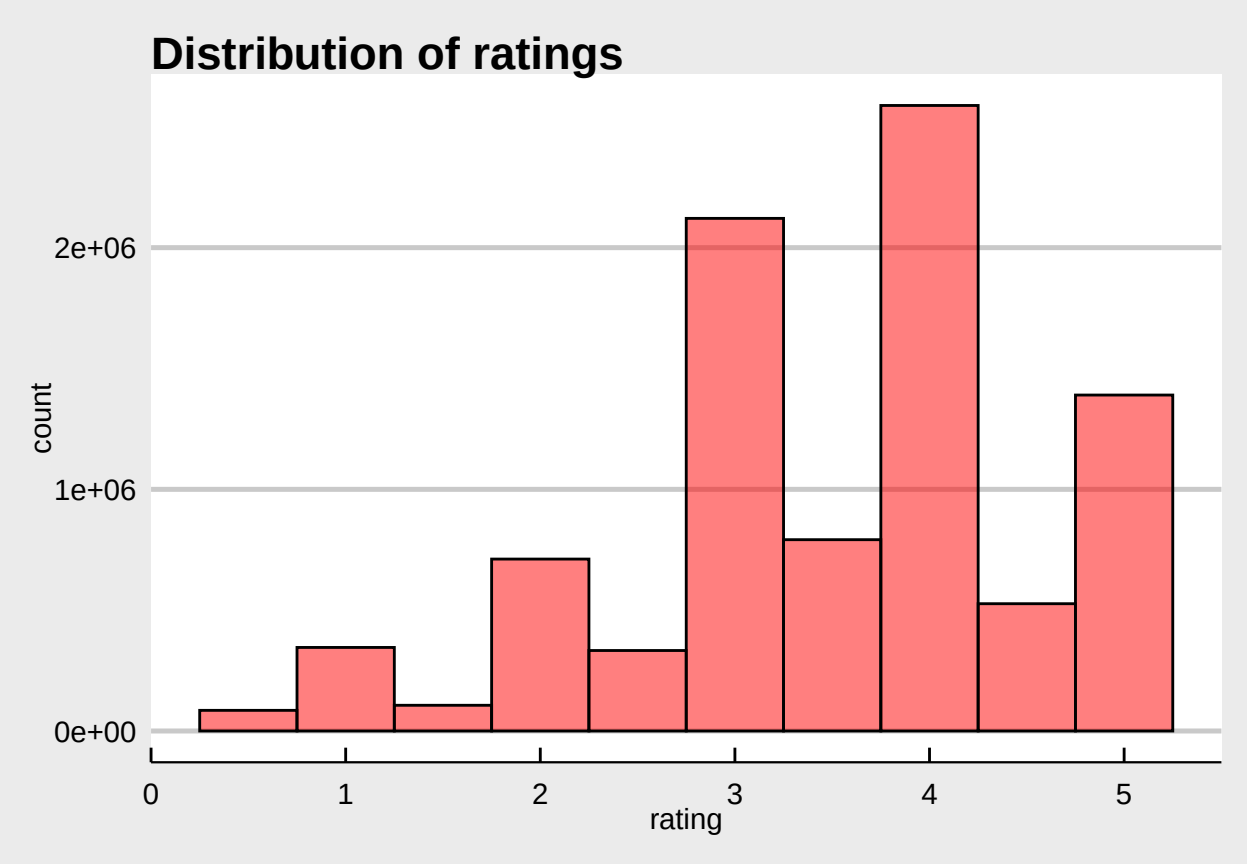
Exploratory Data Analysis

The original 10M dataset is already cleaned from Non Assigned values

Ratings go from 0.5 to 5 with a step of 0.5 (but will be considered as continuous for predictions) and are given by nearly 70,000 users on over 10,000 movies.

n_users	n_movies	min_rating	max_rating
69878	10677	0.5	5

The ratings distribution looks like this :



We can see that most of the ratings are above 2.5. Another striking fact is the affinity for round numbers (nearly 80%). Indeed all the picks we observe appear at whole numbers from 1 to 5.

In order to reduce our analyses' computation time we will start from now to work on a subset of 11% of our data. This lets us 1,000,000 ratings at our disposal which remains a very fair amount to conduct analyses and train models.

3. Model fitting and results

Model picking

In order to build recommandation systems we first need to split our data in training and testing sets. We therefore make a 80/20 split which results in 800,000 observations for training and 200,000 for testing.

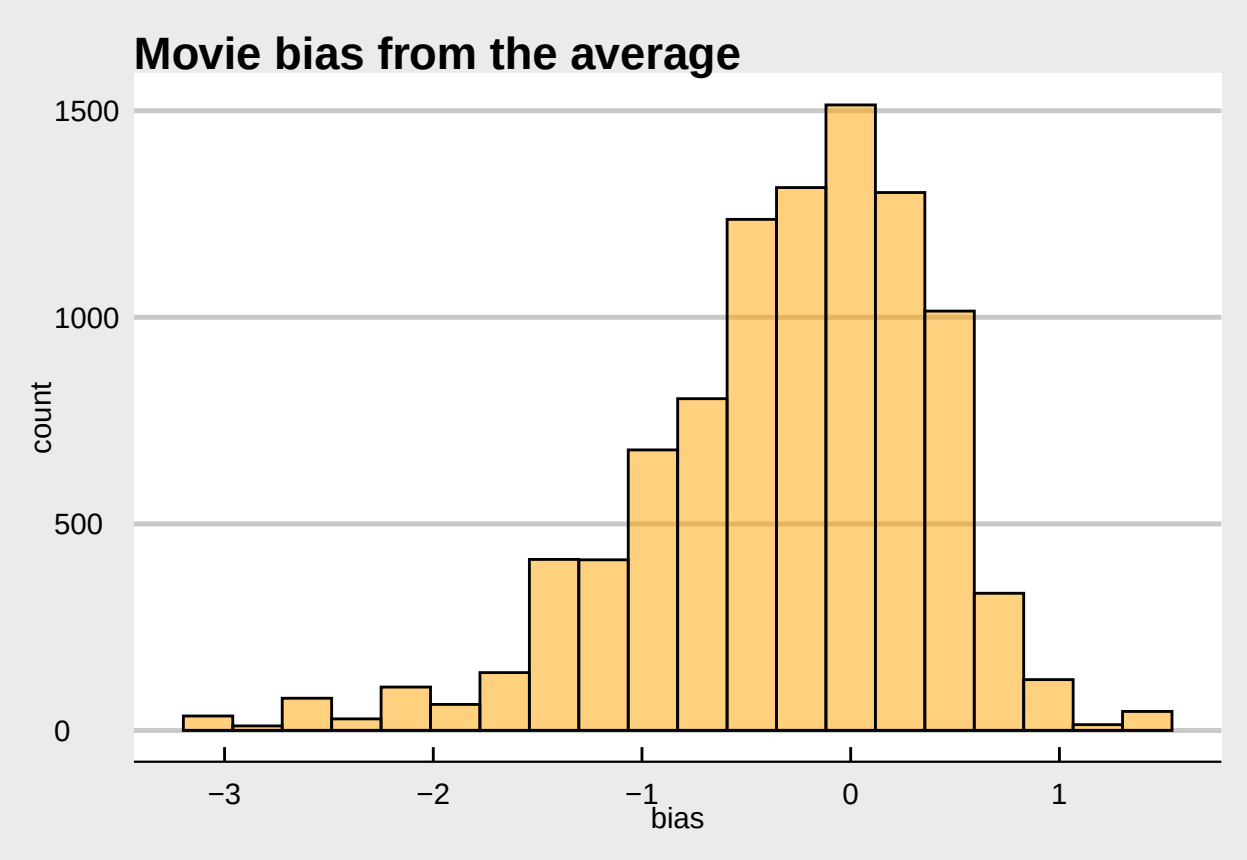
To build our model we start from the basic idea of predicting the same rating for all movies. In this case the forecasted rating would be the average among all ratings we have at our disposal. As we can expect, such a naive model isn't accurate at all. Indeed it leads to an RMSE above 1 which means that on average our predictions diverge by more than one point from the actual ratings. Our goal then is to reduce substantially this divergence.

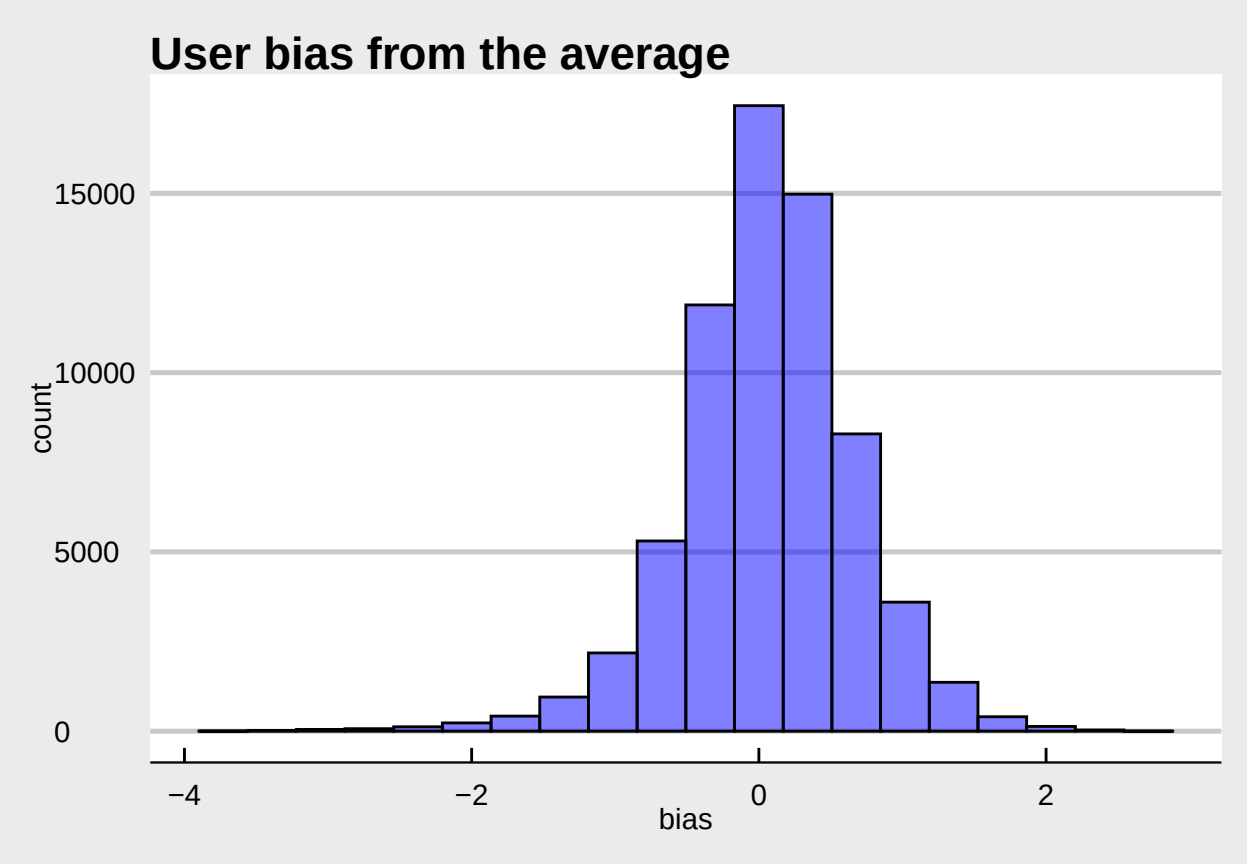
Model	RMSE
Model 0 : Overall Average Rating	1.0613

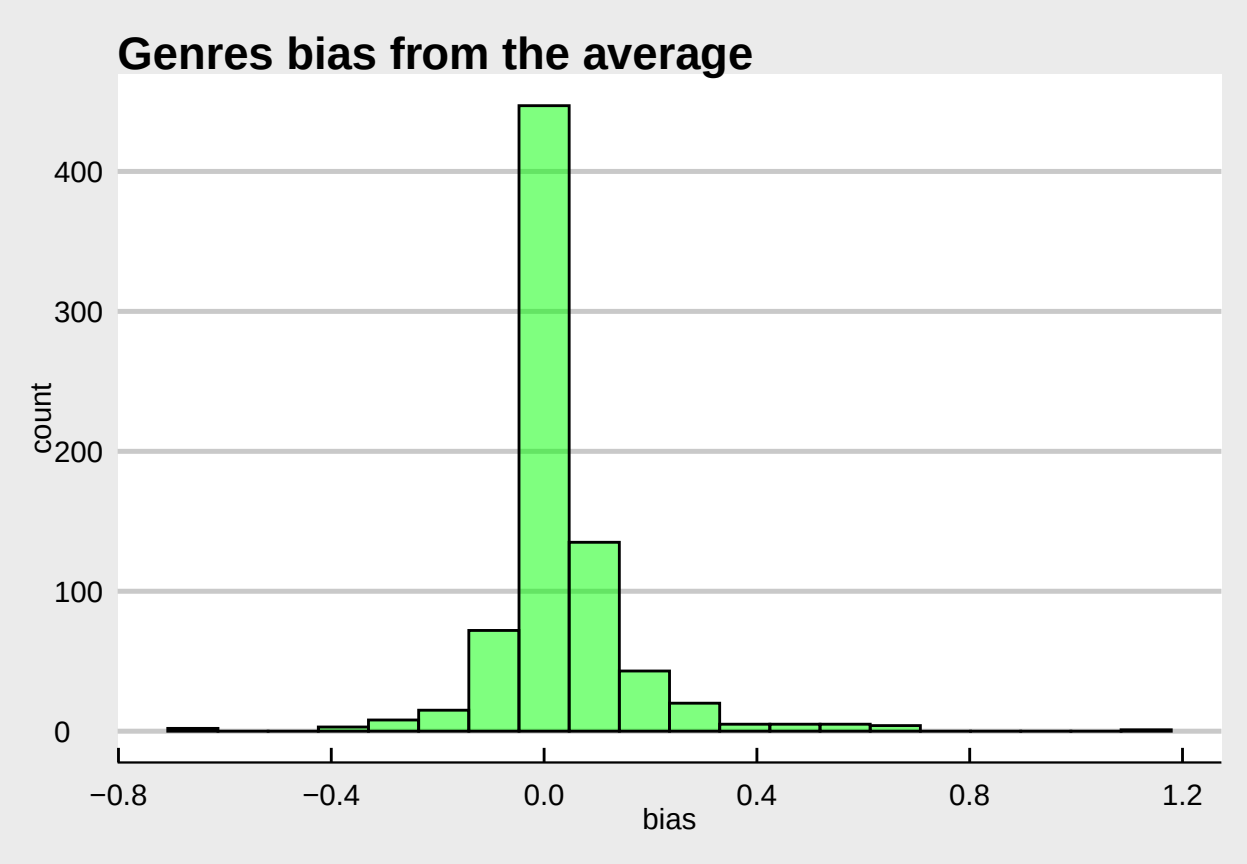
To improve our model we will incorporate what we call biases. For several societal and behavioral reasons, some movies are generally better rated than others while users are more or less critical. As well we have to take into account that users may have preferences for some movie genres.

The following charts quantify these biases and reveal that these are substantial.

- The movie effect is mostly negative.
- The user effect is more heterogeneous
- The genre effect is less important and won't improve the model much




User bias from the average
15000
10000
5000
0
count
−4
−2
0
2
bias



The next table presents the results of these models :

Model	RMSE
Model 0 : Overall Average Rating	1.06125
Model 1 : LM with movie bias	0.94914
Model 2 : Model 1 + user bias	0.91360
Model 3 : Model 2 + genre bias	0.91320

Although these models improve the performance of our predictions they present a critical weakness. The highest estimated ratings are attributed to unpopular movies since we don't have much ratings to predict an accurate estimate. This pattern also happens in the worst estimates. Therefore we need to shrink the lowest and highest estimates so that these unpopular movies converge to the average of similar movies.

To achieve such a correction we use a penalized linear model. Instead of calculating a basic average of the biases, we add a fixed penalty term to the number of ratings per movies to the denominator term of the average formula. As a result this penalty term won't affect much movies that have a large amount of ratings while it will strongly shrink unpopular movie ratings.

We repeat the same approach for the user and the genre bias. The penalty terms were tuned on the training set and we found 2.2 for movies, 5.3 for users and 1.7 for genres.

Model	RMSE
Model 0 : Overall Average Rating	1.06125
Model 1 : LM with movie bias	0.94914
Model 2 : Model 1 + user bias	0.91360
Model 3 : Model 2 + genre bias	0.91320
Model 4 : Regularized Model 1	0.94789
Model 5 : Regularized Model 2	0.88800
Model 6 : Regularized Model 3	0.88773

Now that we have our final model we are ready to test it on the validation set

Final test on the validation set

We get a final RMSE of 0.86368 on the validation set which represents a 18.6% improvement from what we got with a naive model on the training set (Model 0).

Model	RMSE
Model 0 : Overall Average Rating	1.06125
Model 1 : LM with movie bias	0.94914
Model 2 : Model 1 + user bias	0.91360
Model 3 : Model 2 + genre bias	0.91320
Model 4 : Regularized Model 1	0.94789
Model 5 : Regularized Model 2	0.88800
Model 6 : Regularized Model 3	0.88773
Final Model performance	0.86368

4. Conclusion

Our analysis showed that movie and user biases are really important to take into consideration in the model in order to improve our recommendations performance. The genre effect is negligible.

However an RMSE of 0.86368 remains high as it means that on a five rating scale we are on average 0.86 points away from the actual rating. This translates a very low accuracy.

There exist better models today like the Slope One Model that reduces the RMSE to nearly 0.2.

We can note that working on such large data sets is somewhat complicated since it requires high requirements, especially of RAM, in order to compute on all the data. That's why we splitted the data. However it is possible to operate dimensionality reduction techniques and work on more data. Such pre-processing methods are good if the data is appropriate i.e. we don't loe much information doing so.