Practical Machine Learning and Deep Learning

Assignment 2

Andrey Starodumov

B20-DS-01

1 Introduction

The goal of the assignment is to create a model that produce the movie recommendation functionality based on the Movielens dataset which includes the users and movies and the ratings of the films given by users.

2 Data analysis

The data provided to us was explored. And here I want to present the distributions of some features to show why I decided to use the common mothod to create a RecSys model.

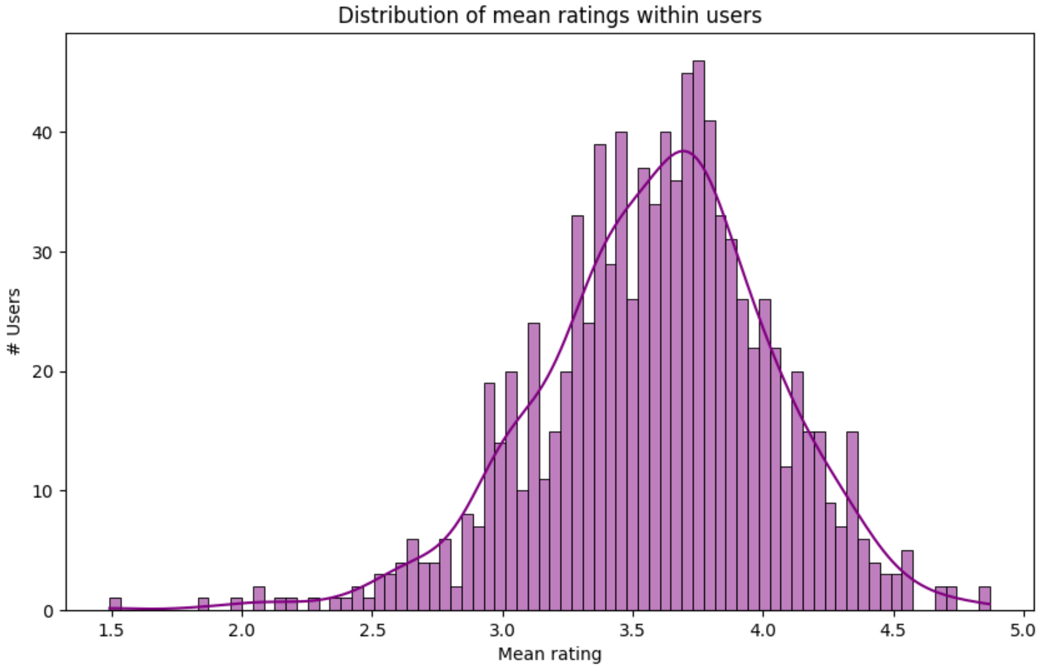


Figure 1: Mean rate of a user on a platform

The main decision was made based on the data of [Fig. 1]. Because of an approximation of the distribution of the mean rated given by each users (out of 943 overall), we can conclude that users rate the movies in a normal manner without huge noise. That is why the method do not uses the gemographic metadata of a users.

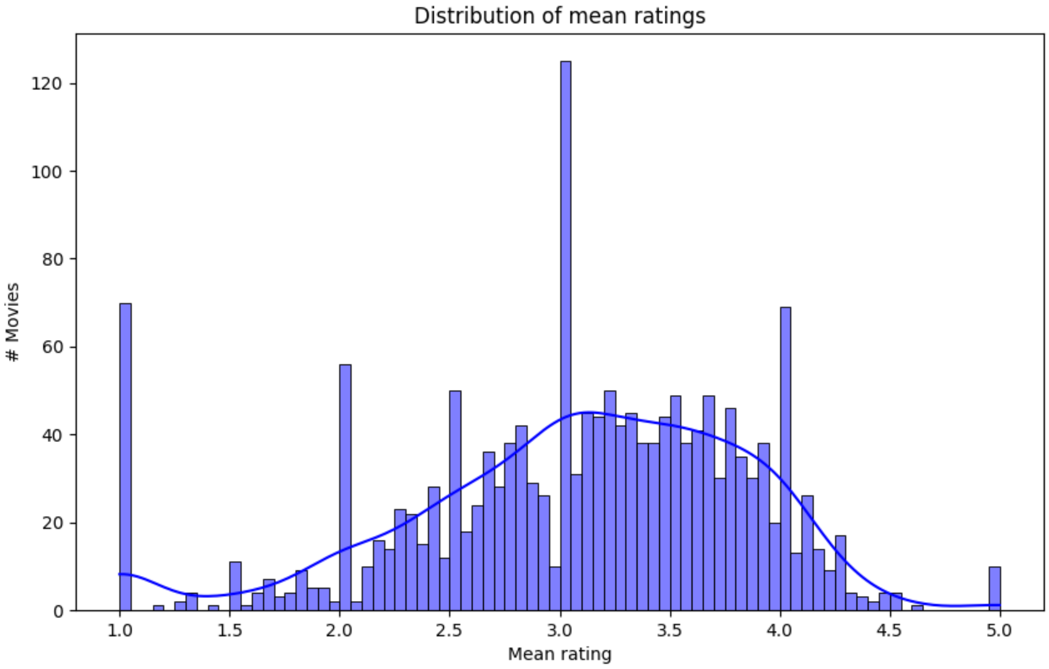


Figure 2: Mean rate of a movie

[Fig. 2] shows that the mean ratings of the movies are also closer to the normal distribution (of course, very dirty approximation). It says that the ratings are quite fair for this platform. And we can easily use all of them. Even those, whose rating is the lowest, because it happens that movies are very bad for some people and they are only a few percent from all movies.

Also I want to point that the number of ratings per users [Fig. 3] is actually a log-normal distribution, which is the show that we deal with the data that is real, because a lot of processes are log-normal.

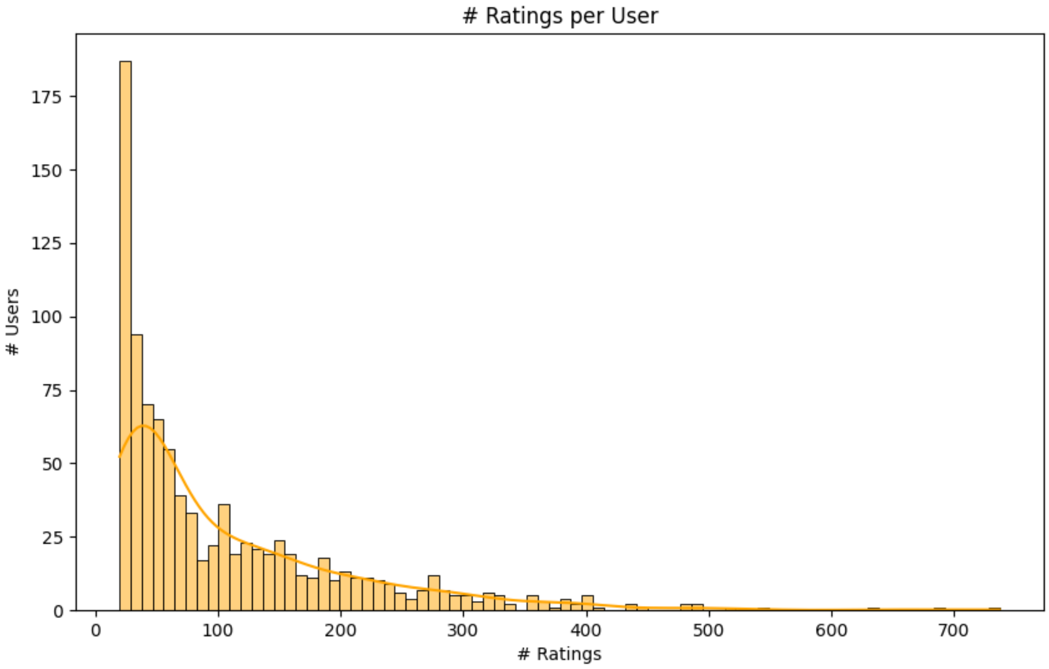


Figure 3: Ratings per User

3 Model Implementation

The model I used is a simple Singular Value Decomposition (SVD). It is a nice and common way to implement recommendation systems using Matrix Factorization algorithm. That is why all that needed for the model is modules of the surprise library to read the data, transform it into the specific class and put into the SVD model via $.fit()$ function.

4 Model Advantages and Disadvantages

This section gives some points about strong and weak sides of the model implemented.

4.1 Advantages:

1. **Personalising:** The model gives recommendations bases on the user preferences which is presented as the ratings of the movies.
2. **Efficiency:** The matrix representation is very efficient way to process enormous amount of the data.

4.2 Disadvantages:

1. **Staring point:** The model gives the same result for the users about whom model does not have an information. And because of the different users' preferences the uniqueness of the starting recommendations is bad.
2. **Context:** the model does not use the additional information, e.g., gemographic information about users, which could be a good for the start point problem.

5 Training Process

The training process here is poor from the RecSys developer side, because all what I use in train the model is $.fit()$ function which is built in model.

6 Evaluation

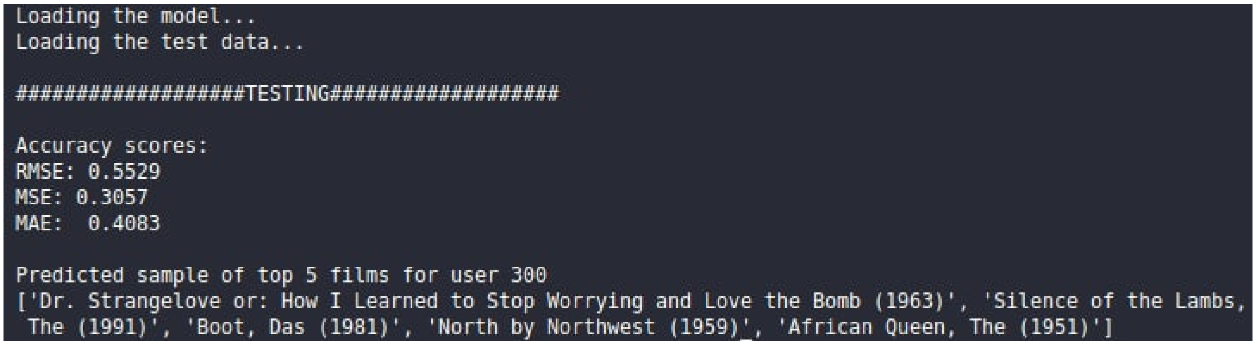
I decided to use several metric for the evaluation of the model on the evaluation data.

Metrics used are:

- **Mean Squared Error:** this metric helps to see how well model performs via punishing it with square function for bad predictions.
- **Root Mean Squared Error:** this metric helps to see how good is model with MSE one.
- **Mean absolute value:** metric shows how good or bad model in sense of bad decisions. Here we assume that the far prediction from true value, the more it means for the bad decision.

All this metrics give more or less clear picture about the model performances. They are used together because of the small values of the regression target (possible rating of the user for the movie).

And since we deal with small values, it will be enough for us if the model will achieve the value form range $[0, 1)$.



```
Loading the model...
Loading the test data...

###################TESTING###################

Accuracy scores:
RMSE: 0.5529
MSE: 0.3057
MAE:  0.4083

Predicted sample of top 5 films for user 300
['Dr. Strangelove or: How I Learned to Stop Worrying and Love the Bomb (1963)', 'Silence of the Lambs,
 The (1991)', 'Boot, Das (1981)', 'North by Northwest (1959)', 'African Queen, The (1951)']
```

Figure 4: Evaluation

7 Results

The model achieved the following errors:

- $MSE = 0.31$;
- $RMSE = 0.55$;
- $MAE = 0.4$.

The result are good for us, and we can use the model as the recommendation system, since all this values are in the range $[0, 0.6)$ which is quite better than $[0, 1)$.

[Fig. 4] show the example of the recommendations for the specific users as well as overall evaluation of the model.