HarvardX Data Science Capstone
MovieLens Project

Pietro D'Ambrosio

October 23, 2019

1. Executive Summary

This project has been implemented as part of the "HarvardX Data Science"" program. The objective of the project is the realization of a system of recommendations starting from a series of 10 million reviews already made by users. After a descriptive analysis of the data and the relative pre-processing necessary to prepare the subsequent processing phases, we proceeded to develop and compare various predictive models. The model that gave the best results was the one based on "matrix factorization". This model was then applied to the validation data set and obtain a RMSE value equal to **0.7908**.

2. Introduction

This section describes the objectives of the project, the input dataset and the key steps that were performed to achieve the results presented in this report.

2.1. Goal of the project

The main objective of the project is to create a system of film recommendations using all the tools learned during the courses and the 10 M version of the MovieLens dataset. The rating forecast will be carried out using automatic learning algorithms using the input data set (edx) and validated using the validation dataset (validation). RMSE was used to assess how close the forecasts were to the actual values of the validation set.

2.2. Input data

The datasets used for this project have been prepared by GroupLens, a research lab in the Department of Computer Science and Engineering at the University of Minnesota Twin Cities specializing in recommender systems, online communities, mobile and ubiquitous technologies, digital libraries, and local geographic information systems. The MovieLens dataset contains 10 million ratings and 100,000 tags applications applied to 10,000 movies by 72,000 users.It was released in January 2009 and is a stable benchmark dataset.

The dataset are publicly downloadable from here.

2.3 Key steps performed

After a first check of the data, a descriptive analysis was carried out by means of research and specific visualizations of datasets in order to obtain useful insights and define the best processing approach.

Subsequently we proceeded to the realization of various modeling hypotheses starting from the simplest possible and increasing their level of complexity. For each modeling hypothesis, a verification of the RMSE was performed and all the values were summarized in a table to allow a more immediate comparison of the results.

3. Methods and analysis

This section explains the process and techniques used, such as data cleaning, data exploration and visualization, the various insights gained, and the modeling approach chosen.

3.1. Process and techniques

The work begins with a quantitative analysis of the acquired data in order to understand the property of the available features and evaluate the need for any pre-processing activity.

Then a data preparation and data cleaning phase is carried out to proceed with the subsequent modeling phases. Finally, the data visualization phase allows the analysis of the single available features and the identification of the most suitable processing strategies to solve the problem.

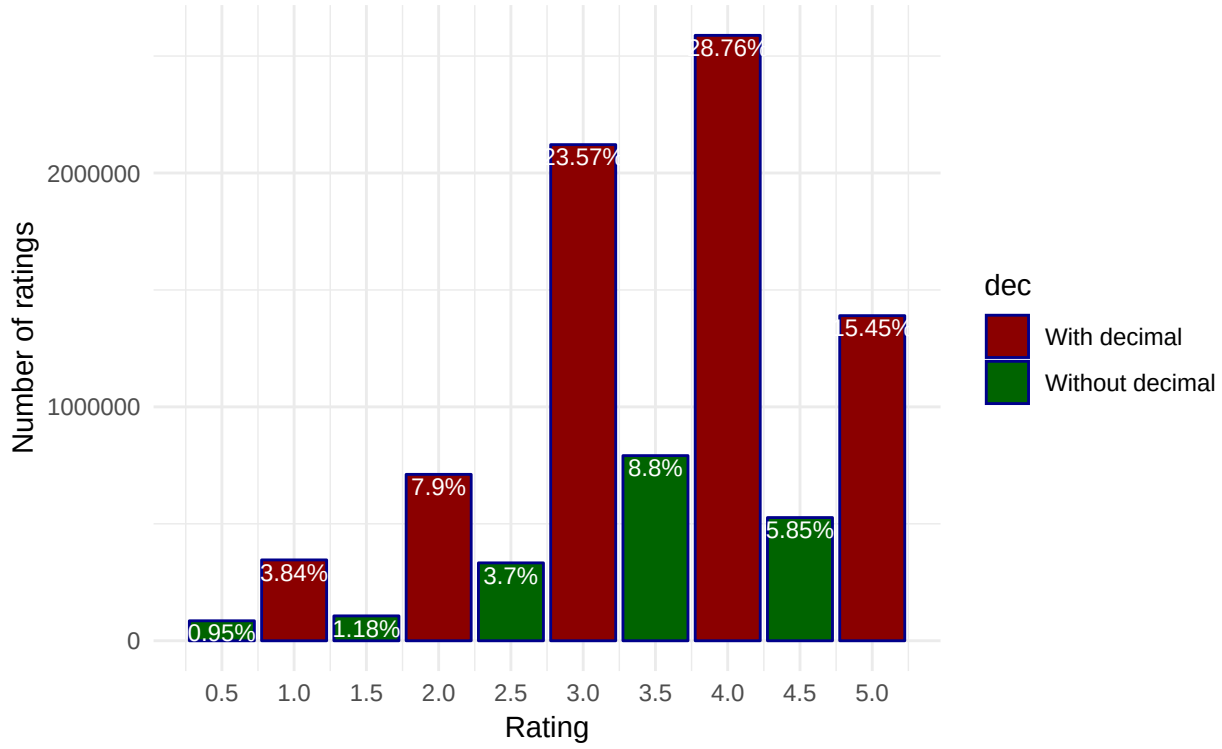
3.1.1. Descriptive analysis

The "edx" dataset contains **9,000,055** rows and **6** variables. The "validation" dataset, instead, contains **999,999** rows. From now on we will only consider the edx dataset since we will use the validation dataset only for the final verification of the results obtained. The following table shows the first lines of the dataset.

	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
Now l	et's exam	ine the co	ntents of	every singl	e variable in the e	dx file:

Field "rating"

This field has a value that varies from 0.5 to 5 (in steps of 0.5) has **mean = 3.51** and **sd = 1.06** and its values are distributed as follows:



From the previous information two elements of interest immediately emerge:

1. the ratings are not uniformly distributed and probably a good part of the users tend to evaluate only the films that they liked

2. ratings with integer values are much more numerous than decimal ratings.

Fields "userId", "movieId" and "timestamp"

The fields "userId" and "movieId" allow us to uniquely identify the users and films to which the assessments relate. The field "timestamp" indicates the date and time the evaluation was performed. In the "edx" dataset there are **69,878** unique users and **10,677** unique movies. So not all users have evaluated all the movies and not all movies are rated by all users.

Field "title"

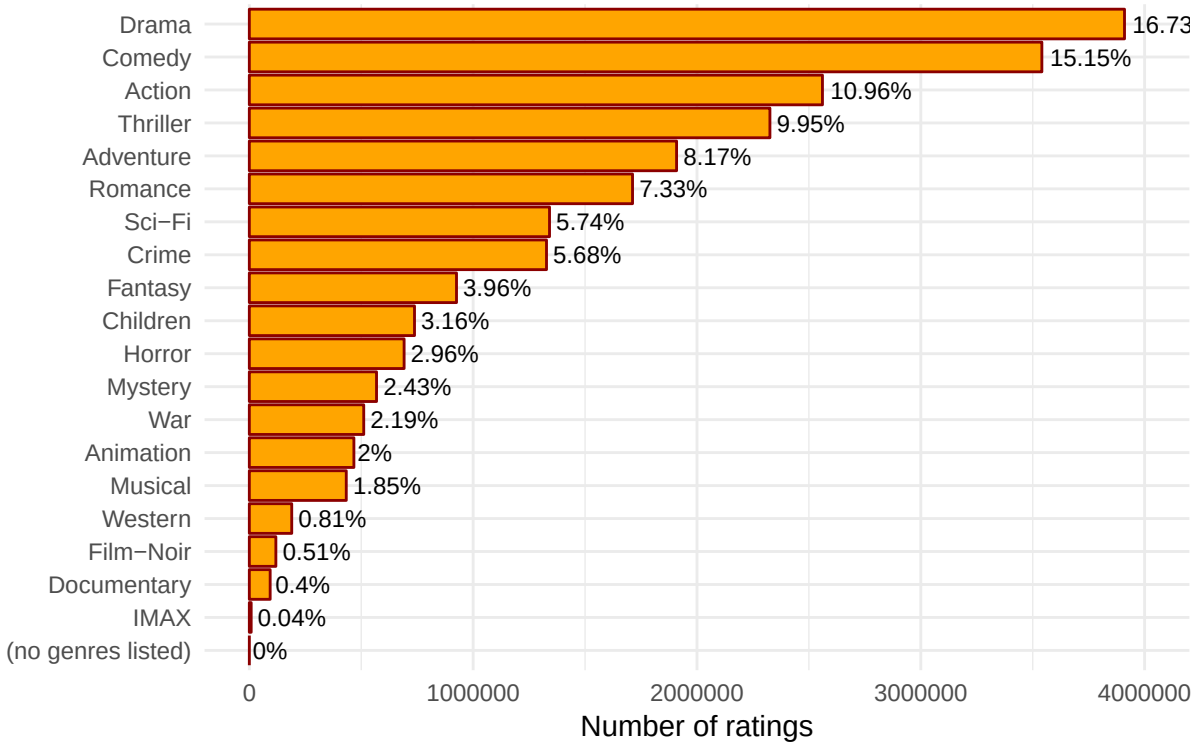
Contains the title and year of release of the film. The year is contained in brackets and is found in the final part of the text. It is certainly advisable to extract this information and keep it in a separate column to be able to use it in the elaborations.

Field "genres"

This field contains the list of genres that characterize the film. If you want to consider the "gender" to be able to carry out evaluations on this information, it is necessary to split this field (and generate separate lines for each gender).

This operation could change the average of the classifications since some films have many associated genres and therefore, after the split, their contribution to the model could be altered.

The following graph shows the frequency of the genres of the films evaluated in descending order:

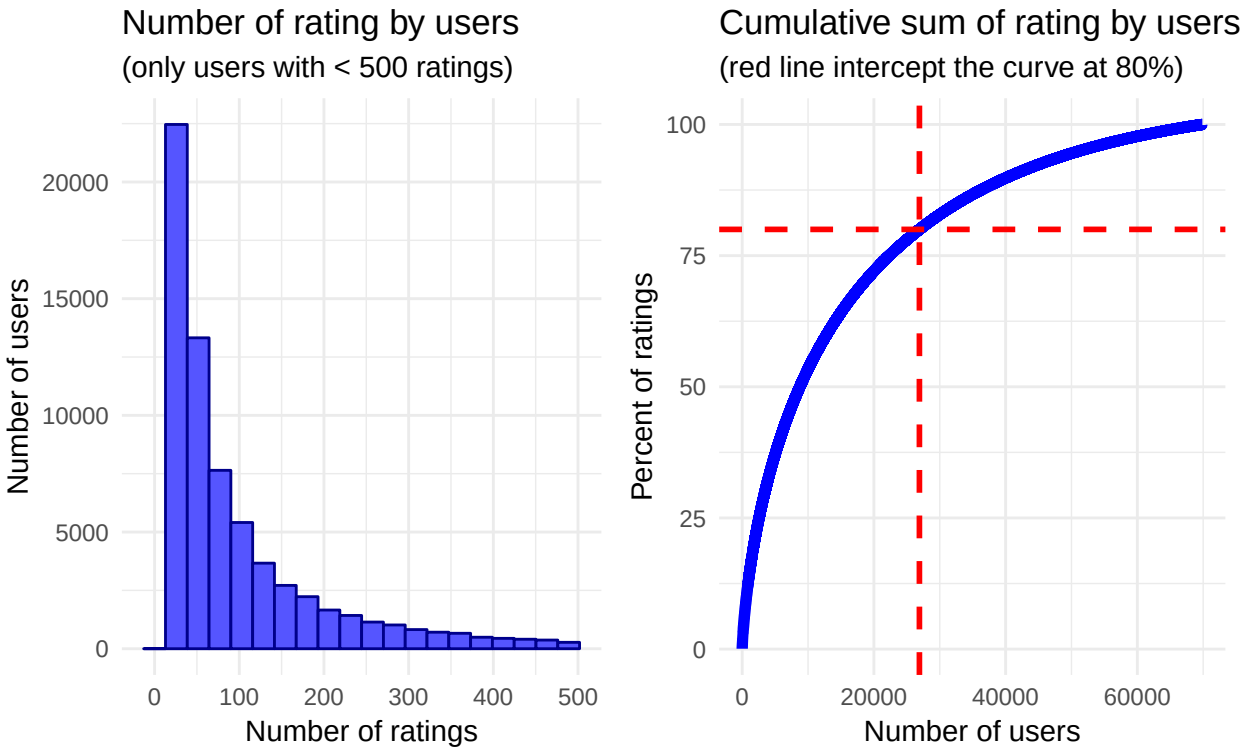


3.1.2. Further data observations

So far we have observed the characteristics of the individual variables. Now let's try to see if there are elements that characterize the "rating" field with respect to the other variables, and in particular: users, films and years.

Rating for Users

The following graphs allow us to see how the number of ratings is distributed for each user. In this case we try to understand if there are users who express more opinions than others and what properties they have (with respect to the opinion expressed).



The previous graphs show that a modest quantity of users (around 38.51) covers 80% of the total quantity of the evaluations carried out. Therefore it is possible that the judgment of some users weighs more than that of other users in the evaluation of the movies.

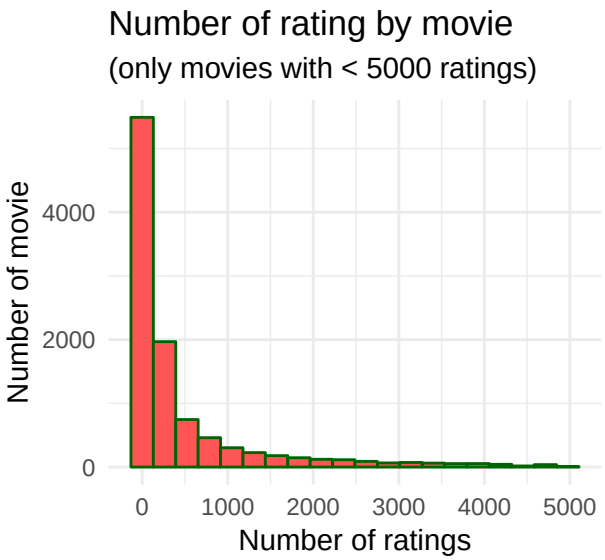
In fact if we observe the average rating obtained on this reduced sample of users we will have a value of **3.4781**, significantly different from that of other users which is **3.6497** (the overall average rating is 3.5125). If we consider an even smaller sample of 15,000 users we will see an even lower average rating (3.4401).

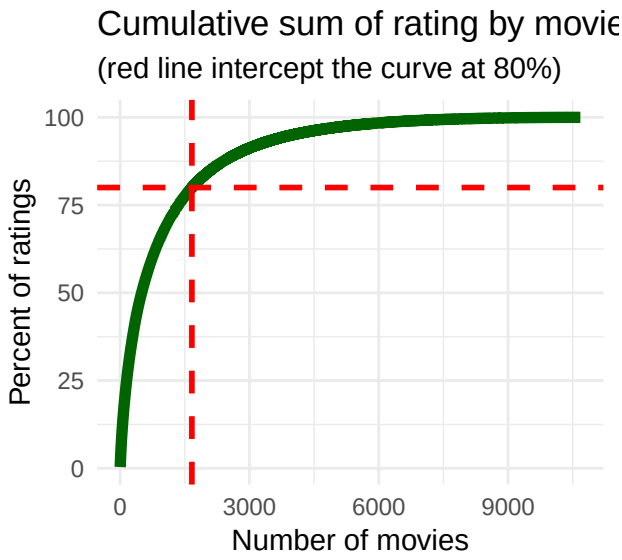
We can therefore deduce that there is an element, which we could define "user factor", which in some way conditions the judgment of a particular user and which can influence the value of the assigned rating.

Rating for Movies

We repeat the analysis just made for users even for movies with the same criteria.

In this case we try to understand if there are movies that tend to have lower or higher scores than the average to see if there is also a "movie factor" to be taken into account in the preparation of the predictive model.





the graphs show that a modest quantity of movie (around 15.58) covers 80% of the total quantity of the evaluations carried out. Therefore it is possible that the mean rating of these movies is different from that of the other movies and this is an important fact for the evaluation model.

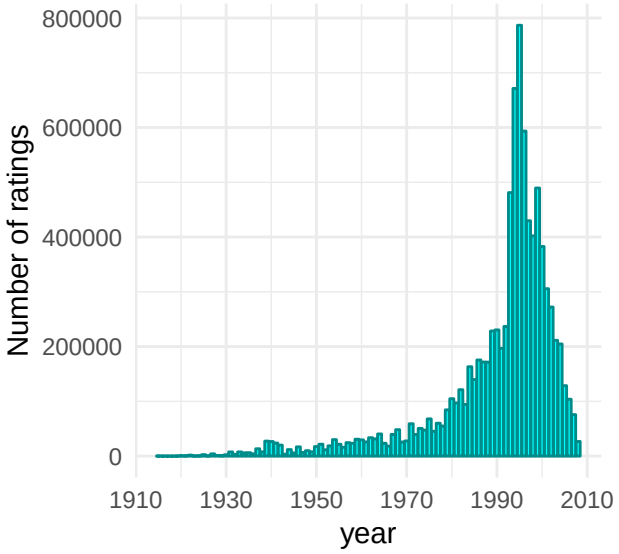
Particularly we can observe the average rating obtained by this subset of movies have a value of **3.5835**, significantly different from that of remain movies which is **3.2283** (the overall average rating is 3.5125). If we consider an even smaller sample of 1,000 movies we will see an even lower average rating (3.6194).

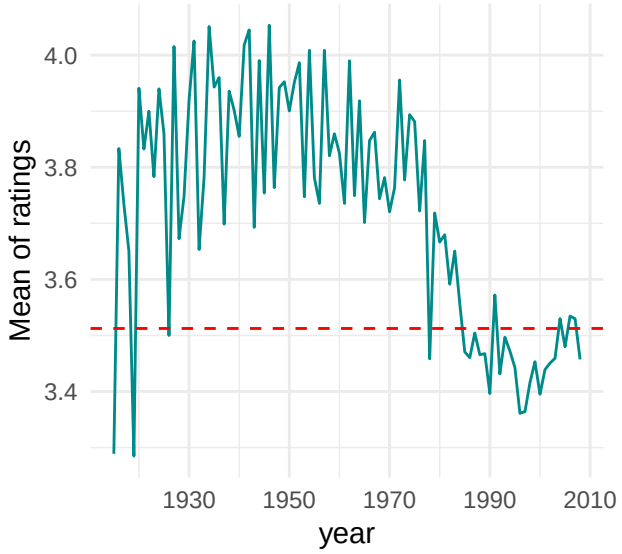
We can therefore deduce that exists also a "movie factor", which in some way conditions the judgment of user for a particular movie and which can influence the value of the assigned rating.

Rating for Year

Let us now try to understand if there is also a "year" characterization of the average ratings assigned by the users.

For this analysis we will use the same criteria that we applied for users and movies but other types of graphics more suitable to represent these characteristics.





It is clear that there is also a "year effect" since the average rating of all movies of a specific year is in some cases significantly different from the overall average. We will therefore consider this factor in defining the strategy of approach to modeling.

3.1.3. Data Pre-processing

During the data observation phase, we found that there are no absent data or other inconsistencies that would force us to perform a data cleaning phase. However, we need to perform some pre-processing on the input data in order to better operate during the modeling phase.

In particular we will extract from the "title" field the year of the film and we will create a new version of the input file with the "genres" field splitted. in this way we can take into account the "genre effect" during the selection phase of the predictive models. Obviously these processes will be performed both on the "edx" dataset and on the "validation" dataset.

After pre-processing the dataset edx appears as follows:

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

3.2. Modeling approach

For choicing the best modeling approach we will proceed step by step, first trying out the regression models (starting from the simplest) and then using matrix factorization which seems to be one of the best approaches for recommendation systems.

Furthermore, to reduce the risk of overfitting, we will first refine the model on a "test" dataset (10% of data extracted from the edx dataset) and then, only at the end, once we have chosen the final model, we will check it on the "validation" dataset. This approach guarantees us greater model reliability in the use on real data.

Below we will present the individual models evaluated, starting from the simplest possible (the simple average of the rating) and continuing with the addition of factors relating to users, movies, genres and years. In some cases we used "cross validation" for the optimization of the model, while for the evaluation of the model we used the typical error loss function RMSE (residual mean squared error), as defined by the following formula:

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}$$

3.2.1. First basic model

This model is as simple as possible and obviously serves as a reference point for the analysis of the performance of the various models we will build and to understand whether we are proceeding in the right way or not.

If our prediction of a rating made by a particular user for a particular movie were based solely on the average of all past forecasts, we would get an RMSE of **1.0601** (the mean of rating). This for us represents the initial value that we should try to improve by adding complexity to the model.

The formula that characterizes this model (that assumes the same rating for all movies and users with all the differences explained by random variation ϵ)is the following:

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

From the evaluation carried out on the test dataset this model presents an RMSE of **1.0601**.

3.2.2. Add movie effect to the model

Now let's see how we can add some elements to the model to consider "the effect of the film" (the average of the historical evaluations obtained by the film) as emerged from the previous considerations.

In this case we can then add to the previous formula a factor that depends on the previous evaluations obtained by the movie. The new model can be formally described as follows:

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

In this case the term

$$b_i$$

represents average ranking for the movie "i".

Once the factor b_i has been calculated for each movie, we merge this information in the "train" dataset and then train the model and apply it to the "test" dataset. the RMSE calculated with this model is equal to **0.943**.

This RMSE is better than the previous one, as shown in the summary table shown below.

method	RMSE
Model 1 - Simple average of rating	1.0601
Model 2 - Add movie effect	0.943

3.2.3. Add user effect to the model

In this step we add new element to the model to consider the so-called "user effect" (the average of the historical evaluations given by the user) as emerged from the previous considerations.

So we can add to the previous formula a factor that depends on all the evaluations given by the specific user. The new model can be formally described as follows:

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

The term b_u represents average ranking for the user "u".

Once the factor **b_u** has been calculated for each user, we merge it in the "train" dataset, train the model and then apply it to the "test" dataset. the RMSE calculated with this new model is equal to **0.8647**.

We have achieved a further improvement of the RMSE compared to the previous models, as shown in the following summary table.

method	RMSE
Model 1 - Simple average of rating	1.0601
Model 2 - Add movie effect	0.943
Model 3 - Add user effect	0.8647

3.2.4. Add year effect to the model

Now we add further element to the model to consider the "year effect" (the average of the historical evaluations obtained by all the movie of a specific year) as emerged from the previous considerations.

In this case we can add to the previous formula a factor that depends on the previous evaluations obtained by all the movie of a specific year. The new model can be formally described as follows:

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

The term b_y represents average ranking for the year "y".

The factor b_y has been calculated for each year and merged in the "train" dataset. The model applied to the "test" dataset gave a RMSE equal to **0.8643**.

The obtained RMSE shows that the model guarantees better performance than the previous ones.

method	RMSE
Model 1 - Simple average of rating	1.0601
Model 2 - Add movie effect	0.943
Model 3 - Add user effect	0.8647
Model 4 - Add year effect	0.8643

3.2.5. Add genre effect to the model

Finally consider the "genre effect" (the average of the historical evaluations obtained by all the movies of a specific genre) as emerged from the previous considerations.

In order to perform this processing we need to split the "genres" field to separate and weigh the various genres associated with the single movie separately.

In this case we add to the formula a factor that depends on the previous evaluations obtained by all the movie of a specific genre. The new model can be formally described as follows:

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

The term b_g represents average ranking for the genre "g".

Once the factor b_g has been calculated for each genre, we merge it in the "train" dataset, train the model and apply it to the "test" dataset.

The RMSE calculated with this model is equal to **0.8624**. Below is the summary table of the results obtained.

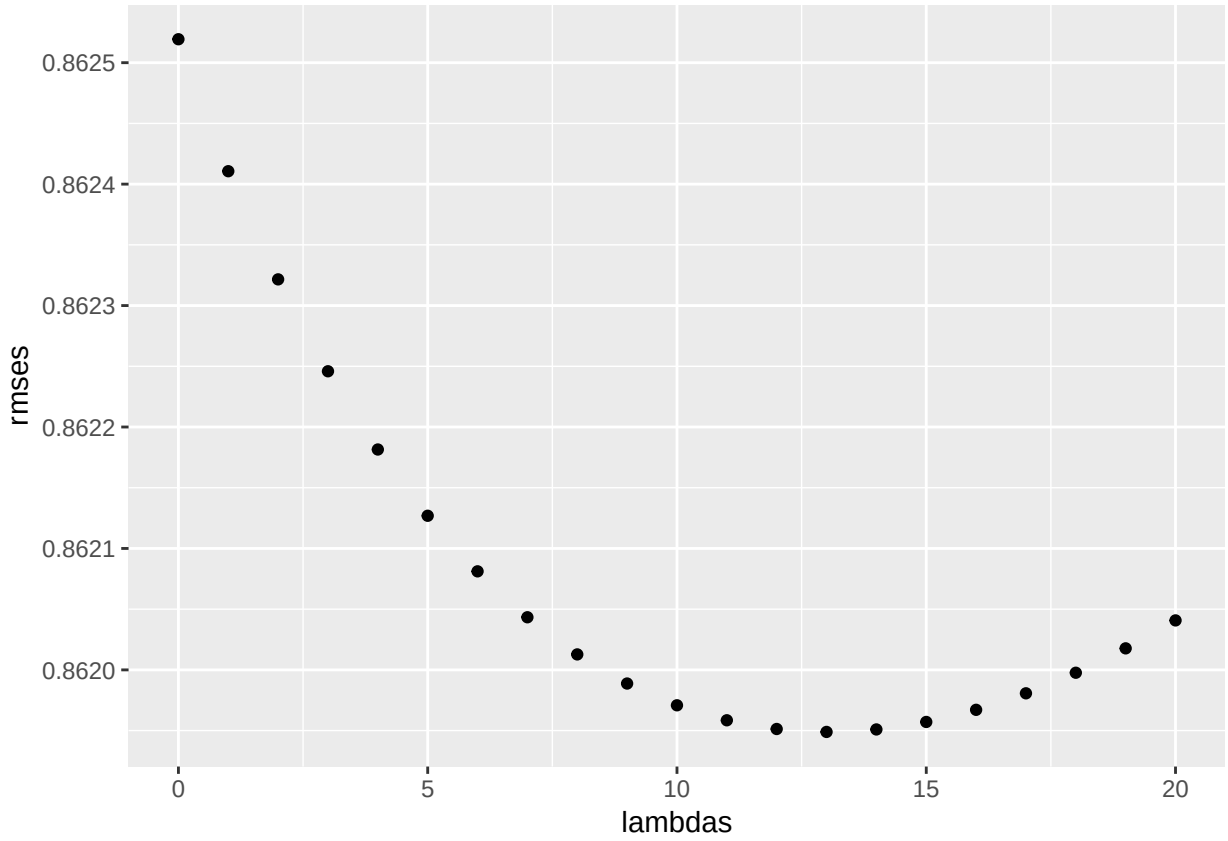
method	RMSE
Model 1 - Simple average of rating	1.0601
Model 2 - Add movie effect	0.943
Model 3 - Add user effect	0.8647
Model 4 - Add year effect	0.8643
Model 5 - Add genre effect	0.8624

3.2.6. Model regularization

The previous model could be further improved by applying the "regularization" to penalize the contribution to the model of the factors calculated on very small groups (eg Users who have evaluated only one film or films evaluated by a few users).

This penalty factor, which we will call Lambda, is made in such a way that when the number of observations is high it becomes practically zero, while it is significant if the number of observations is low.

The tuning of the lambda penalty factor was evaluated with values ranging from 0 to 20 (step 1). As shown in the following graph, the best value was found to be **13**.



Applying the regularization to the model the obtained RMSE is equal to **0.8619**. Below is the summary table of the results obtained.

method	RMSE
Model 1 - Simple average of rating	1.0601
Model 2 - Add movie effect	0.943
Model 3 - Add user effect	0.8647
Model 4 - Add year effect	0.8643
Model 5 - Add genre effect	0.8624
Model 6 - Add regularization	0.8619

3.2.7. Matrix factorization

The last model has not produced substantial improvements to the results obtained from the previous models, so we will examine another approach this time based on matrix factorization.

This technique allows us to consider a new source of variation in values, namely the fact that some films tend to be judged in a similar way to others (for example sagas, sequels, etc.) and some users tend to express opinions similar to other users.

Furthermore, matrix factorization is widely used in machine learning in recommendation systems because it produces very effective results in this type of problem.

To proceed with this approach it is necessary to prepare a matrix in which all the users are the rows and all the films are the columns and represent in each cell of the matrix the residuals, thus defined:

$$r_{u,i} = r_{u,i}$$

The size of the matrix, the complexity of the calculations and the need to carry out various parameters have led to very high processing times. Upon completion of the processing, the RMSE value found on the test data set is equal to **0.7938**. This value is much better than all those obtained with the other models based on regression.

If we consider that the range of the ratings varies between 0.5 and 5 and that among the predicted values there are numerous values that lie outside this range, we can apply a simple heuristic to force the predictions out of range to the limit values 0.5 and 5.

Applying this further adjustment we get a final RMSE equal to **0.7935**.

4. Results

4.1. modeling results

The following table shows the results of all the models, verified on the test dataset.

method	RMSE
Model 1 - Simple average of rating	1.0601
Model 2 - Add movie effect	0.943
Model 3 - Add user effect	0.8647
Model 4 - Add year effect	0.8643
Model 5 - Add genre effect	0.8624
Model 6 - Add regularization	0.8619
Model 7 - matrix Factorization	0.7938
Model 8 - matrix Factorization with range control	0.7935

The best model is undoubtedly the one based on matrix factorization. We will then use this model, with the same parameters used previously, on the "validation" dataset to determine the final RMSE.

4.2. modeling performance

The matrix factorization model, verified on the "validation" dataset, recorded a final RMSE of **0.7908**, a result that can certainly be considered interesting considering the performances obtained with the initial models, the small number of variables used and the computer resources available (a common notebook).

5. Conclusion

5.1. brief summary of the report

The work began with a descriptive analysis of the information contained in the input datasets with the aim of obtaining useful insights to guide the subsequent modeling phases.

The approach chosen for modeling was to extract a subset of tests from the train dataset for not to use the validation dataset for the development of the models and reduce the risk of overfitting.

As regards predictive models, models based on linear regression (with increasing complexity) were first tested. Then we proceeded to apply the matrix factorization which showed the best results. The model chosen was then applied to the "validation" set, obtaining a final RMSE equal to **0.7908**.

5.2. limitations and future work

It would be interesting to improve the model taking into account other information that was not taken into consideration at this stage, such as the time elapsed between the first review of the film and the year of

release, or to better analyze the borderline cases in order to identify any additional effects to be taken into consideration.

One could also consider the possibility of making an ensemble of models to try to compensate for the weaknesses of each and get better results.