Capstone: Movielens Project

Gustavo Mayeregger

2020-04-16

Introduction

This project consist in a movie recommendation system for the movielens dataset. One of the most popular implementation of the machine learning are the recommendation systems, they are everywhere, to each search of any kind of products we have a list of products that can interested us and in consequence more are the chances that we buy something.

The movies recommendation become popular thanks to Netflix, who is the pioneer in movies streaming. With this system Netflix can predict which movies are the best recommendation to keep viewing movies.

Evaluation method

The method chosen to generate the prediction is minimize the Residual Mean Square Error (RMSE) between the prediction ($\hat{y}$) en the actual values (y):

$$RMSE = \sqrt{\frac{1}{N}\sum (y - \hat{y})^2}$$

The data exploration

For this project the chosen dataset is the 10 millions movielens dataset, that consist in 10 millions independent ratings. The dataset can be downloaded from: http://files.grouplens.org/datasets/movielens/ml-10m.zip

To train and test purpose the movielens data set was divided in to datasets:

- The edx dataset with the 90% of the ratings. In this dataset is made all the data analysis, visualization and train algorithm.
- And the validation dataset in which are made the prediction and final RMSE.

Both datasets have the following fields:

	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

- userID: the ID of each user.
- movieID: the ID of each movie.
- rating: the rate given to a movie from 0 to 5 by half points.
- timestamp: the time and day of the rate. Is given in seconds and begin the count in 1970-01-01.

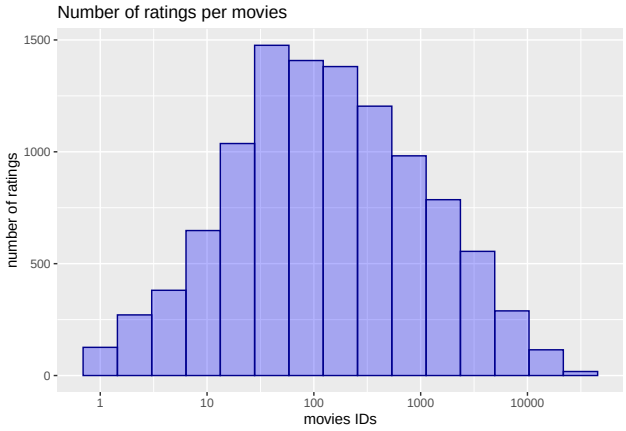
- title: the title of the movie. Also contain the year of the movie release.
- genres: the genre of the movie. One movie can be in multiples genres.

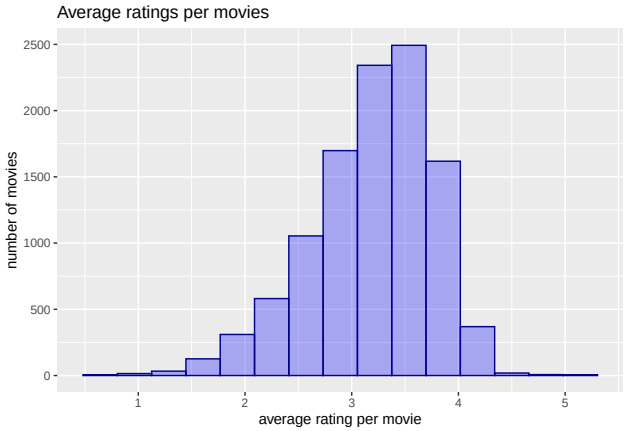
Data Analysis

In this section we will explore the data to find all the variables that effect the ratings. The best way is arrange the data and visualize it through graphics.

Movie effect

There is obvious that movies influence the number of ratings and the rating itself. In the following two graphics we can see that there is movies with few number of ratings and others with a lot of ratings. Also there are movies with a low average rating from the users and others with more average rating.

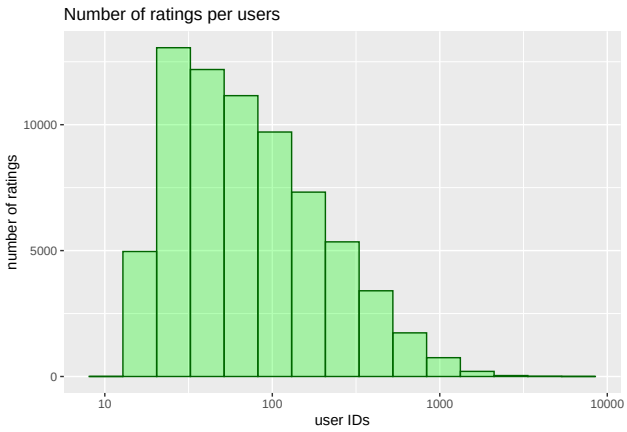


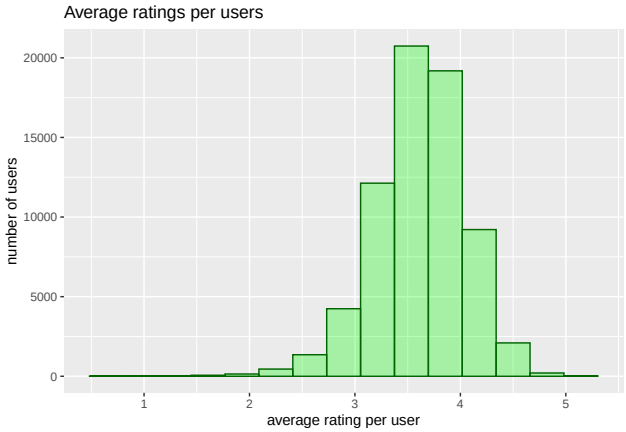


The relation between movies and ratings is very strong and is a variable we need use in the prediction model.

User effect

The same thing that occur with the movies occur with the users. There is users that rate few movies and users that rate a lot of movies. And there is users that rate low and others that place high ratings for the movies.



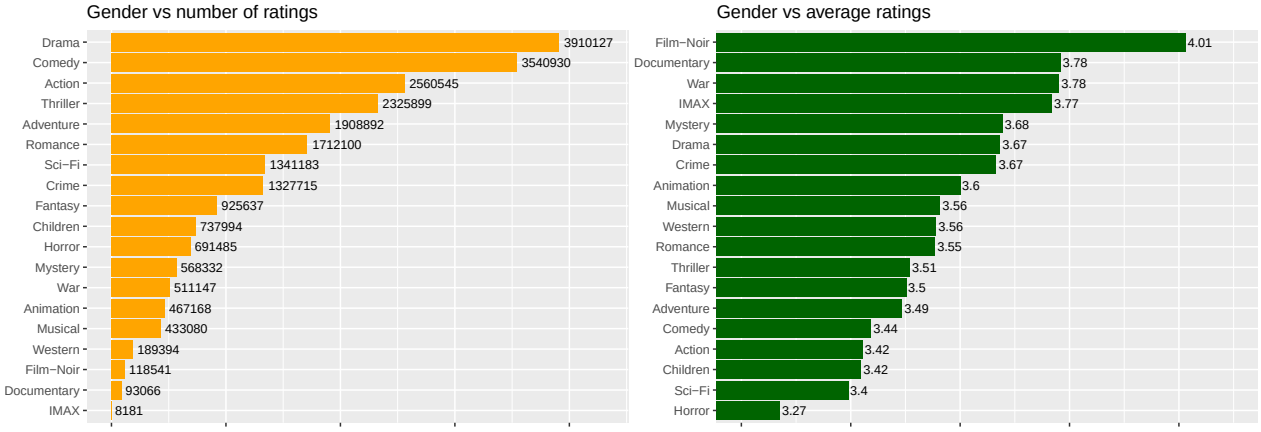


Also this variable will be included for the prediction model.

Gender effect

Individual genres

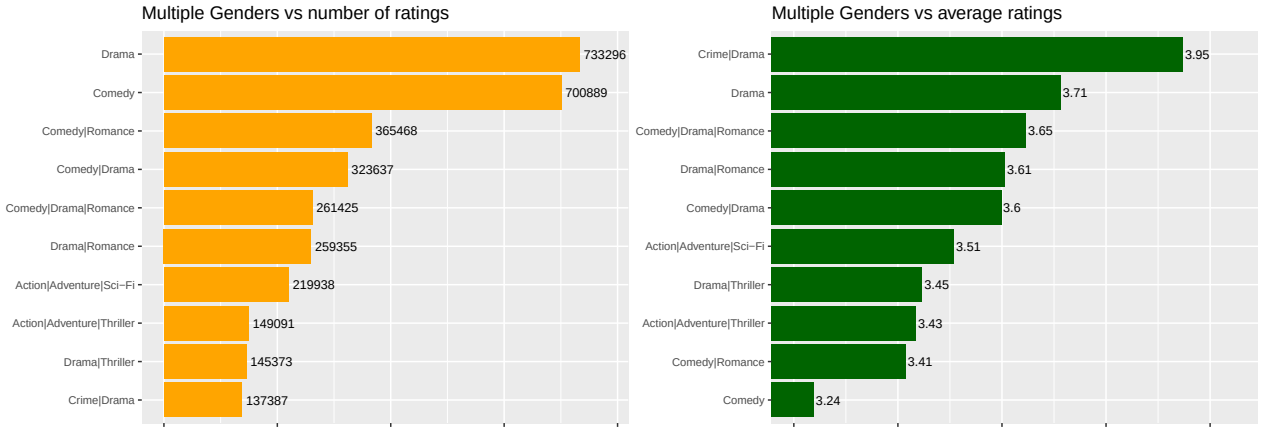
In the dataset there are 19 different genres as can we see here:



The most rated are the dramas, comedys and action movies. But the movies with the best ratings are the film-noir, documentary and war movies as we can see in the next chart.

Multiple genres

But in the movielens dataset the genre of the movies can consists in multiple genres. One movie can be adventure, action and comedy at the same time.



If we look at the multiples genres format also the most rated movies are the dramas, comedy and action genres and the genre with the best rating are the Crime|Drama combination.

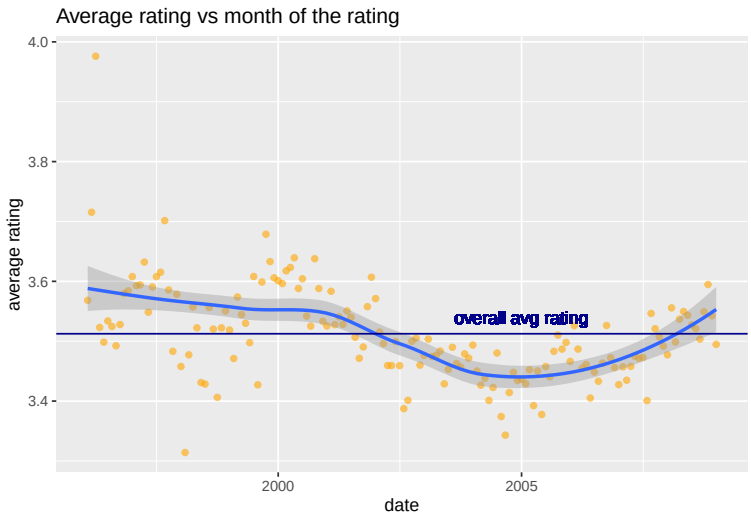
With this analysis we conclude that the genre of a movie influence the rating of the users and must be considered as a variable in the prediction model. The difference if we take the individuals genres or the multiple genders is not significant and we choose the multiple gender variable as the dataset give us.

Timeline effect

Now we going to analyze the behavior of the ratings in the time. To do that we will use the timestamp provided in the dataset to calculate different time lapses like years, days, days of the week and hour of the day.

Ratings through the years

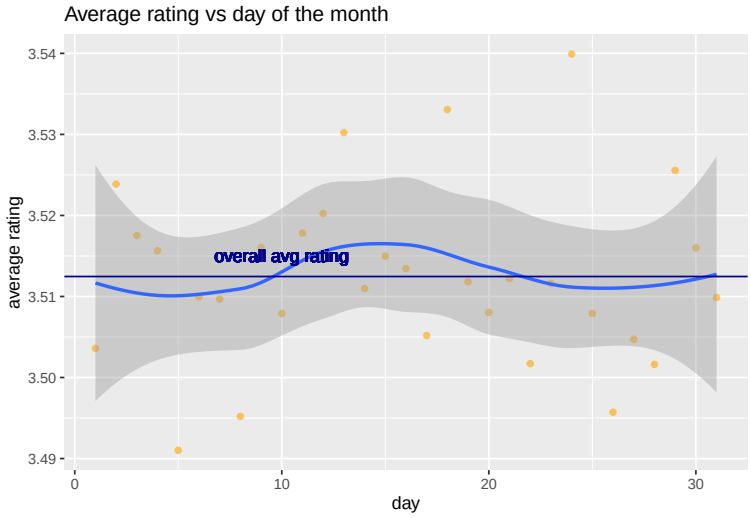
To study this we took the timestamp and separate in all the months from the date of the first rate (january 1995) to the last (january 2009). For each month we calculate the average rating and we can see them in the next graphic.



In the graphic we can see that the rating across the years has been changing comparing to the overall rating of 3.51. From 1995 to 2002 the average rating were higher tan the overall rating. But from 2002 to 2007 the average rating were lower than the overall rating. In the 2 last years the ratings begun increasing again.

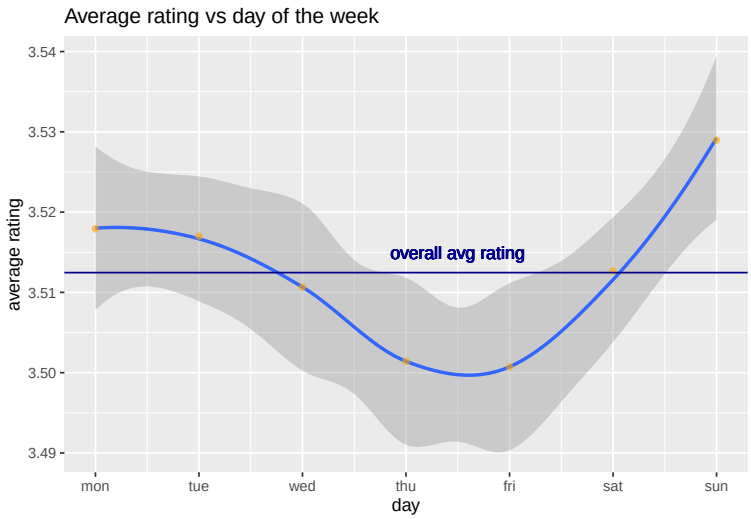
The average ratings variations are between 3.3 to 3.7 that is a significant range and we must include this variable in the prediction model.

Ratings through the days of the month



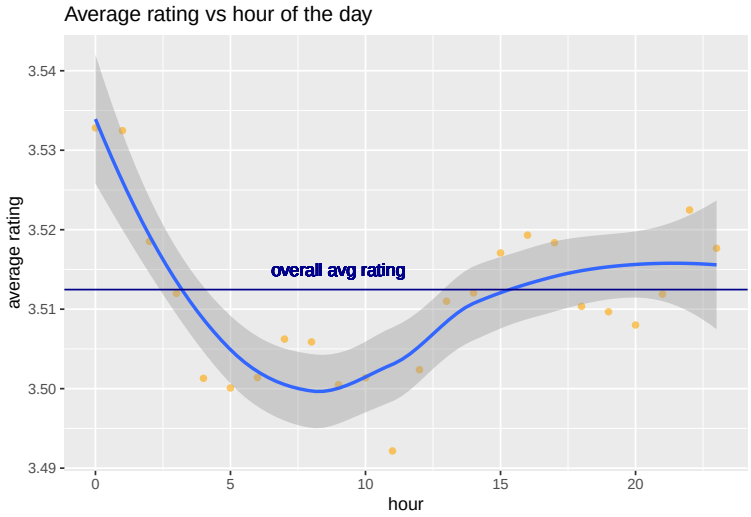
Here we can´t see a clear relation between the day on the month and the ratings and we will not consider this variable.

Ratings through the days of the week



In this case there is a slightly relation between the day of the week and the rating. in the weekends increase the average rating and in the middle of the week decrease. But the change (3.50 to 3.53) is not considerably.

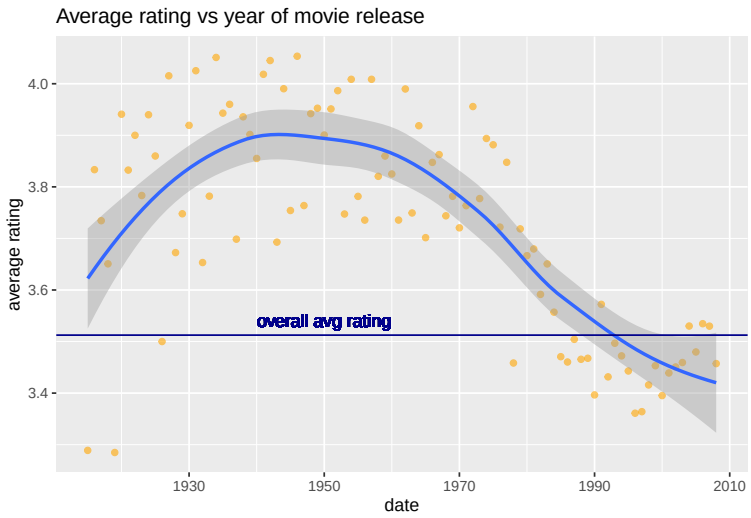
Ratings through the hour of the day



As happened with the days of the week here there is a slightly relation, we can see that the average rating increase at the and of the day and decrease in the morning. But this changes (3.50 to 3.53) are also not considerably for the prediction model.

Year of movie release effect

As we saw in the introduction section, the titles of the movies in the dataset include also the year of the movie release. With this new variable we will analyze the relation between the ratings and this years. Is normal imagine that old movies are for a selected group of users and this users overrating them.



In the graphic we can see a strong relation between the release year of the movie and the ratings. For the movies older than 1990 the average rating increase considerably till averages near 4.0. But for the movies released after 1990 the average rating decrease to 3.4.

This variable have a strong relation with the rating and we must include it in our prediction model.

Modeling method

In the analysis section we conclude that the variables that have more influence in the ratings are: movies, users, genres, date of rating and year of movie release.

The simplest method

The simplest method consist in calculate the overall average rating (μ) and then make all the predictions equal to μ. The resulting RMSE is:

MODEL	RMSE
Just the average	1.061202

Movie and user effect

The changes that appear in the rating from movie to movie and user to user can be modeled like this:

$$\hat{r}_{iu} = \mu + b_i + b_u + \epsilon_{iu}$$

Where:

$\hat{r}_{iu}$ is the predicted rating considering the movies (i) and the users (u)

μ in the overall rating average of the dataset,

b_i is the bias of each movie from μ,

b_u is the bias of each user from μ and

ϵ_{iu} is a random error variable.

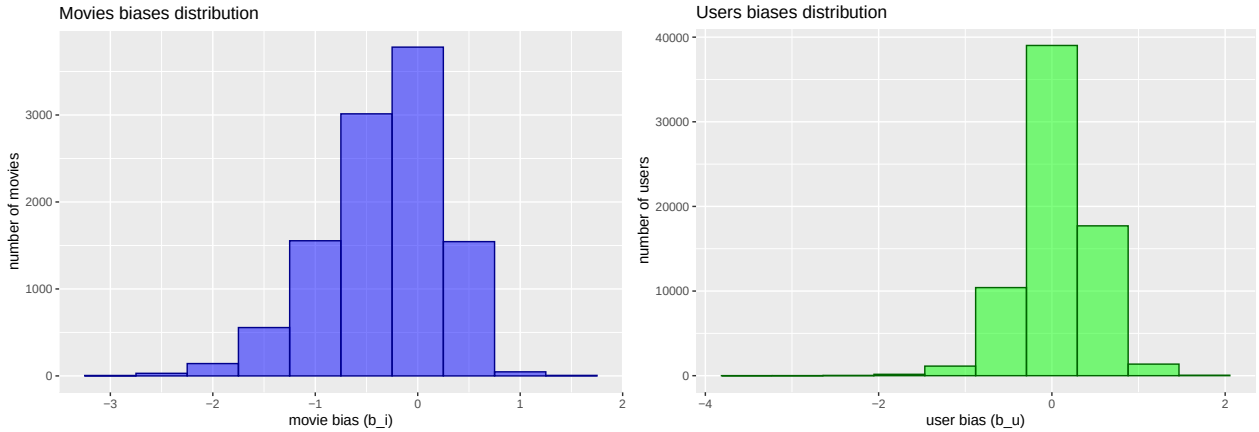
Each bias can be calculated in this way. Note that to calculate the bias of the users we consider the bias of the movies in the mean calculus.

$$b_i = \frac{1}{n_i}\sum_k (r_{i_k} - \mu) \qquad b_u = \frac{1}{n_u}\sum_k (r_{u_k} - b_{i_k} - \mu)$$

Where r_{i_k} are the observed ratings and n_i is the number of ratings for all the movies. In the same way r_{u_k} are the observed ratings and n_u the number of ratings for all the users.

Note that we can use a linear model function to calculate the same but this method consumes much more time for the machine because our extense dataset.

First let´s compute the b_i and b_u bias in the edx dataset. then the prediction and compare it to the validation dataset with the RMSE function.



In this pictures we can see that the biases have a normal distribution, and the variation is between -3.5 and 1.5 as we expect because the μ is 3.5.

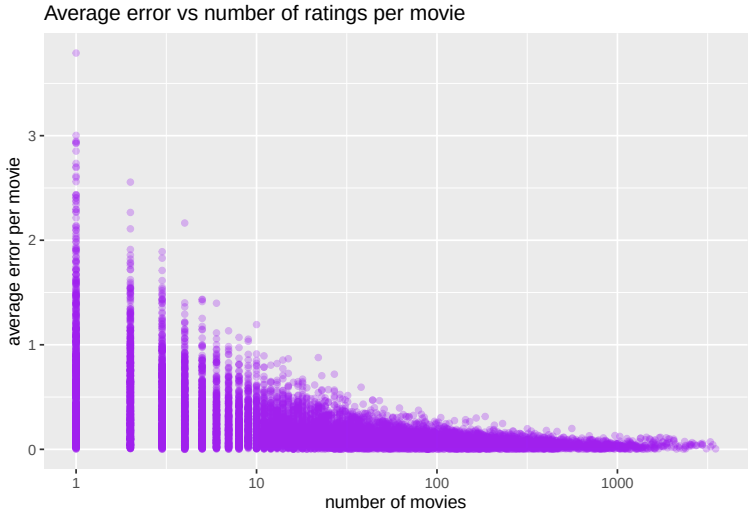
Now we can calculate the ratings predictions with $\hat{r}_{iu} = \mu + b_i + b_u + \epsilon_{iu}$ and then the RMSE with the predicted ratings and the ratings in the validation set.

MODEL	RMSE
Just the average	1.0612018
Movie & User Effect	0.8653488

With the obtained result now we can search the greatest errors obtained by the use of the studied model.

Based in the predictions used to calculate the RSME we can calculate the error in the predictions the difference between the predicted ratings and the validation dataset ratings.

If we average this errors by movies, count the ratings per movies and then plot that is clear that the movies with less number of ratings have the greatest errors.



This problem is explained by the fact that when we calculate the biases b_i and b_u we use the mean function. The more number of ratings we calculate the mean the less is the standard error ($se = \overline{x}/\sqrt{n}$). One method to solve this problem is called regularization.

Movie and user effect with regularization

The regularization method consist in penalize the mean by adding a constant λ to the n

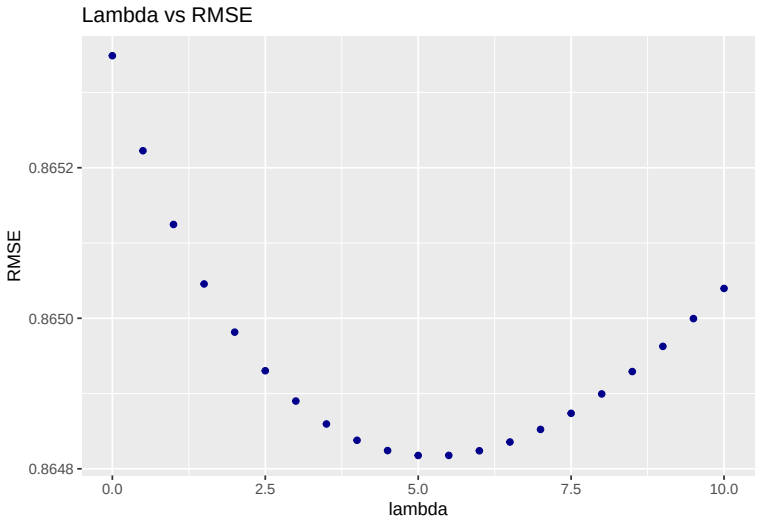
$$\overline{x} = \frac{1}{n}\sum_k x_k \qquad\qquad \overline{x} = \frac{1}{n+\lambda}\sum_k x_k$$

When n is very large compared with λ the effect of λ in the mean is not significant. But when n is equal or less than λ the mean depends of λ. With this we can keep a lower standard error $se = \overline{x}/\sqrt{n+\lambda}$ thanks to the penalization.

Let´s use this method in the previous model to compare the obtained RMSE.

Because the λ is a constant, we need a series of then to see which one produce the minimum RMSE.

$$\lambda = 0, 0.5, 1, 1.5, ..., 10$$



In the graphic we can see that the minimum RMSE correspond to a λ of 5. The regularized model have a best RMSE than the previous one without regularization.

MODEL	RMSE
Just the average	1.0612018
Movie & User Effect	0.8653488
Regularized Movie & User	0.8648177

Movie, user, genre, date & release model with regularization

User and movies are very important variables, but as we saw in the analysis section there are others like the genre, the date of the rating and the year of the movie release.

This five variables generate each one a bias from the overall average rating like we saw earlier for movies (b_i) and users (b_u).

$$b_i = \frac{1}{n_i+\lambda}\sum_k (r_{i_k} - \mu) \qquad\qquad b_u = \frac{1}{n_u+\lambda}\sum_k (r_{u_k} - b_{i_k} - \mu)$$

The other three biases are b_g for genres, b_d for the date of the rating and b_r for the year of the movie release.

$$b_g = \frac{1}{n_g + \lambda} \sum_k (r_{g_k} - b_{i_k} - b_{u_k} - \mu) \qquad b_d = \frac{1}{n_d + \lambda} \sum_k (r_{d_k} - b_{i_k} - b_{u_k} - b_{g_k} - \mu)$$

$$b_r = \frac{1}{n_r + \lambda} \sum_k (r_{r_k} - b_{i_k} - b_{u_k} - b_{g_k} - b_{d_k} - \mu)$$

And the predicted rating with those biases is:

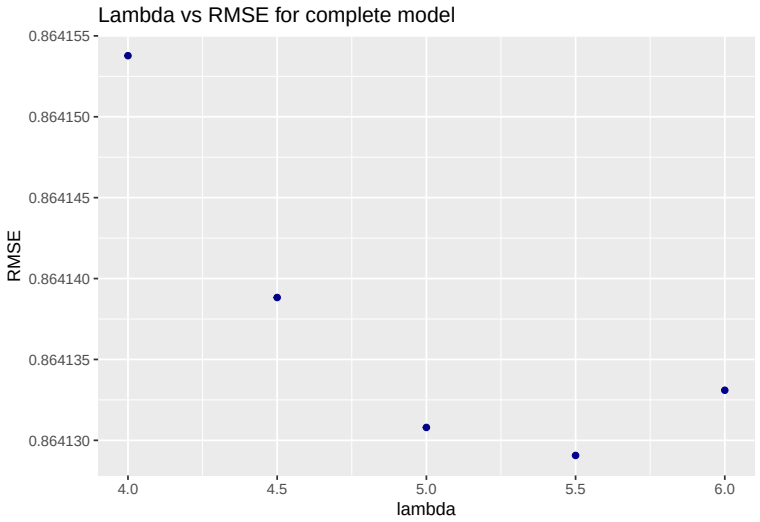
$$\hat{r}_{iugdr} = \mu + b_i + b_u + b_g + b_d + b_r + \epsilon_{iu}$$

Note that in the calculus of the biases we include the penalization constant λ to obtain a regularized model.

Because we aggregate new variables to our model, we need to re calculate the λ for this complete set of variables.

We know that the best λ is near 5 and now we will use a set of values near 5 to find the best tune.

$$\lambda = 4, 4.5, 5.0, 5.5, 6$$



From the graphic we can take that the best λ is 5.5 and the RMSE is even lower than the previous model.

MODEL	RMSE
Just the average	1.0612018
Movie & User Effect	0.8653488
Regularized Movie & User	0.8648177
Regularized Movie, User, Gender, Date & Release	0.8641291

Results

With the simplest method that consist in predict all like the overall average we obtain a RMSE = 1.0612. We will use these value to calculate the improvement of each model vs just the average.

MODEL	RMSE	IMPROVE
Just the average	1.0612018	0.0000000
Movie & User Effect	0.8653488	0.1845577
Regularized Movie & User	0.8648177	0.1850582
Regularized Movie, User, Gender, Date & Release	0.8641291	0.1857071

Table 1: Best Movies without Regularization

title	count
Blue Light, The (Das Blaue Licht) (1932)	1
Class, The (Entre les Murs) (2008)	3
Constantine's Sword (2007)	2
Fighting Elegy (Kenka erejii) (1966)	1
Human Condition II, The (Ningen no joken II) (1959)	4
Human Condition III, The (Ningen no joken III) (1961)	4
Mickey (2003)	1
Satan's Tango (Sátántangó) (1994)	2
Sun Alley (Sonnenallee) (1999)	1
Who's Singin' Over There? (a.k.a. Who Sings Over There) (Ko to tamo peva) (1980)	4

Table 2: Best Movies with Regularization

title	count
Casablanca (1942)	11232
Dark Knight, The (2008)	2353
Godfather, The (1972)	17747
Godfather: Part II, The (1974)	11920
Mickey (2003)	1
Rear Window (1954)	7935
Schindler's List (1993)	23193
Seven Samurai (Shichinin no samurai) (1954)	5190
Shawshank Redemption, The (1994)	28015
Usual Suspects, The (1995)	21648

There is a 18.45% reduction in the RMSE using the Movie & User effect model. This change is the biggest as we can expect because in a movie recommendation system the two principal components are the movies and users. All the others variables will produce smallest improvements.

With the regularized model we improve the previous model for smallest number of ratings. That change we can see if we compare the best and worst movies with the number of ratings of each ones.

After look the tables we can see the importance of the regularization. The improve in the precision of the predictions (18.51% compared to the average) is not too much but affect at the best and worst predictions of the dataset that are just the most searched parameters.

The addition of the others variables to the model (genre, date and release) cause a higher improve than the model with only the movies and users (18.57% compared to the average).

Conclusion

The goal of this project was design a movie recommendation system for the Movielens dataset with 10 millions of ratings. One of the mayor complexity is just manage the 10 millions of observations.

We studied various methods (lm, glm, knn, rpart, rf) to model a system like this but based on a smaller dataset (1000000 observations). If we use any of those models for sure we will achieve a smaller RSME but that will take too much time and the risk of crash the software.

For that reason the chosen method was modeling with the bias from the overall rating for each variable to calculate the prediction. This method is not the more effective but is simple and take less time to the computer than the others methods.

Another method used was the regularization during the biases calculus. This technique help to a better prediction when the number of observations (ratings) is very low, to smaller observations we have greater standard errors.

A future work can consist in explore an others methods, like the dimension reduction, principal component analysis and matrix factorization. But first I need to upgrade my computer for those methods because they consumes too much resources.

I´m very happy with all I learned in this 9 courses of HarvardX. Is my first online course and is the first time I face with the data science. I have a grade in mathematics and most of the statistics themes were familiar to me. The mayor challenge was the coding and this course helped to me so much.