HarvardX – Data Science Capstone: MovieLens Capstone Project Report

Marisa Ivonne Zamora Carrillo

Contents

Overview

Recommendations systems or recommender systems use historic information to generate recommendations for the users. Previous information is used to predict what rating that person is going to give to something and then recommend that to the users.

Some of the most famous companies that use recommendation systems area Amazon, Netflix, Spotify and LinkedIn, based on the information they have about the items (jobs, movies, music, clothes, etc.) that you have rated, the predict the rating than you are most likely to give to other items and those are their recommendations.

In this project the goal is to build movie recommendation system that is going to be evaluated based on the **RMSE (Root Mean Squared Error)**

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

The goal is to reach a RMSE lower than **0.86490**

Data Analysis

The data set used in this project is provided by GroupLens, available in this website http://files.grouplens.org/datasets

The 10 million row dataset is divided into two datasets. The training and the validation dataset which is 10 percent of the data.

The training dataset contains 9 million of rows, almost 70,000 users and approximately 10,000 different movies

Rows	Users	Movies
9000055	69878	10677

The variables in both datasets area the following

```
[1] "userId"    "movieId"   "rating"    "timestamp" "title"     "genres"
```

The edx data set looks like this:

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

Wangle and prepare the data

In the data analysis we find out these things of the dataset that we are going to change so that the dataset is ready for the model building phase:

1. Separate the genre
2. Get the date from the time stamp and separate the in year and month column
3. Extract the year of the movie from the title

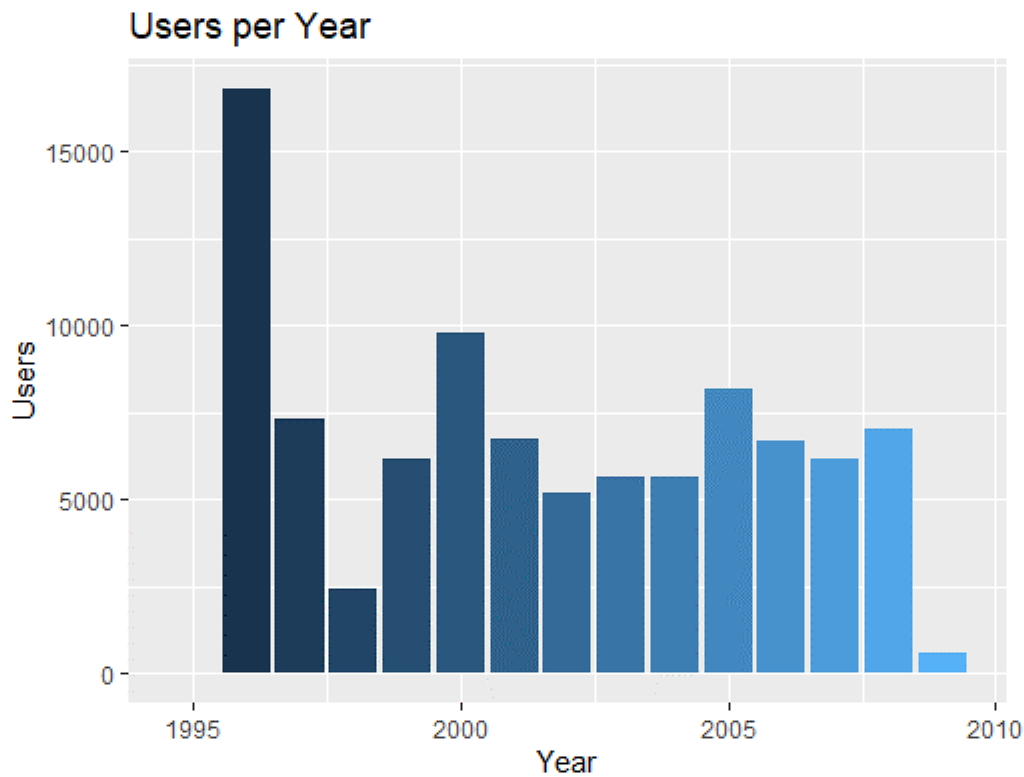
We are going to do this for the edx and the validation dataset. After the data preparation the datasets look like this:

userId	movieId	rating	title	genres	release	month	year
1	122	5	Boomerang (1992)	Comedy	1992	08	1996
1	122	5	Boomerang (1992)	Romance	1992	08	1996
1	185	5	Net, The (1995)	Action	1995	08	1996
1	185	5	Net, The (1995)	Crime	1995	08	1996
1	185	5	Net, The (1995)	Thriller	1995	08	1996
1	292	5	Outbreak (1995)	Action	1995	08	1996

Exploratory Analysis

Users

Users that rated movies per year

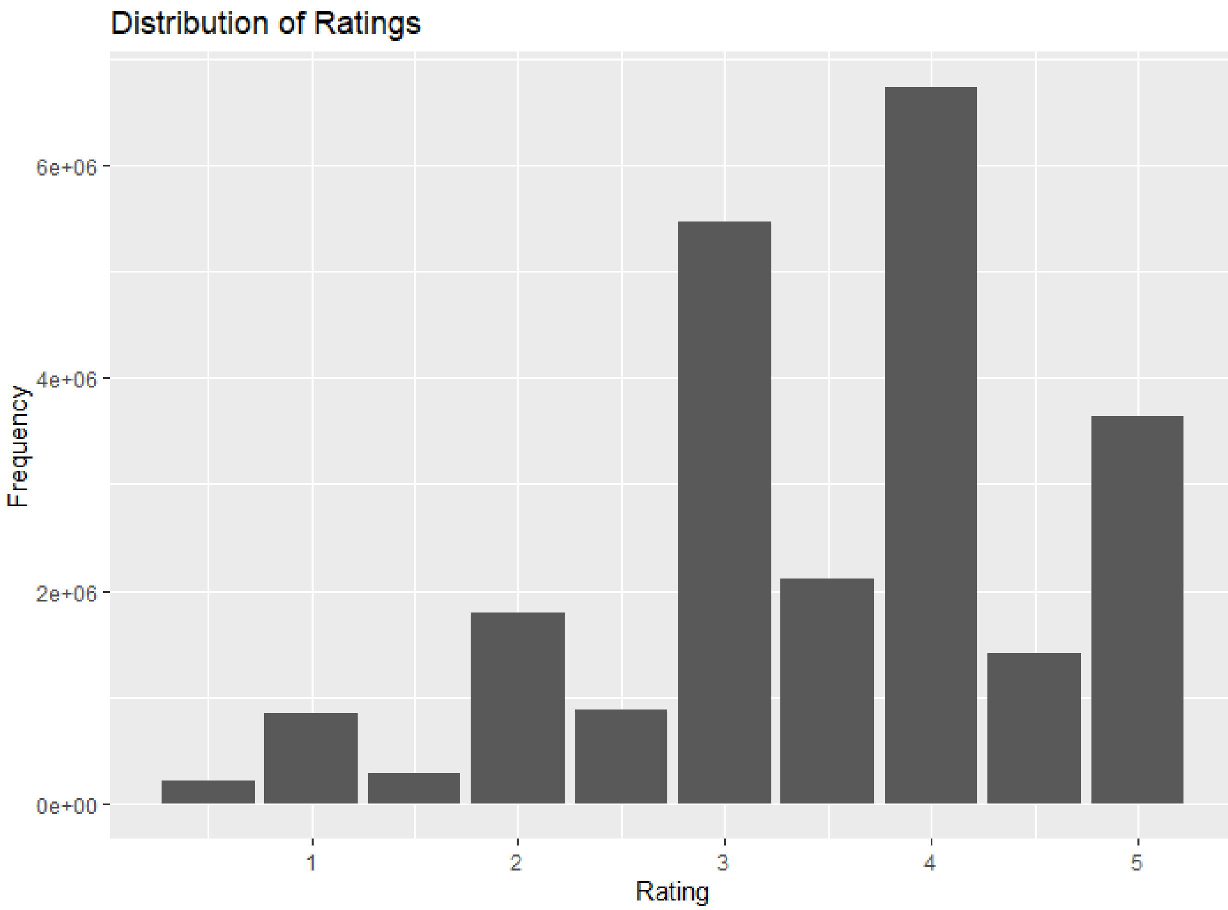


Top 10 ratings per user per year: These are the users that submit more ratings

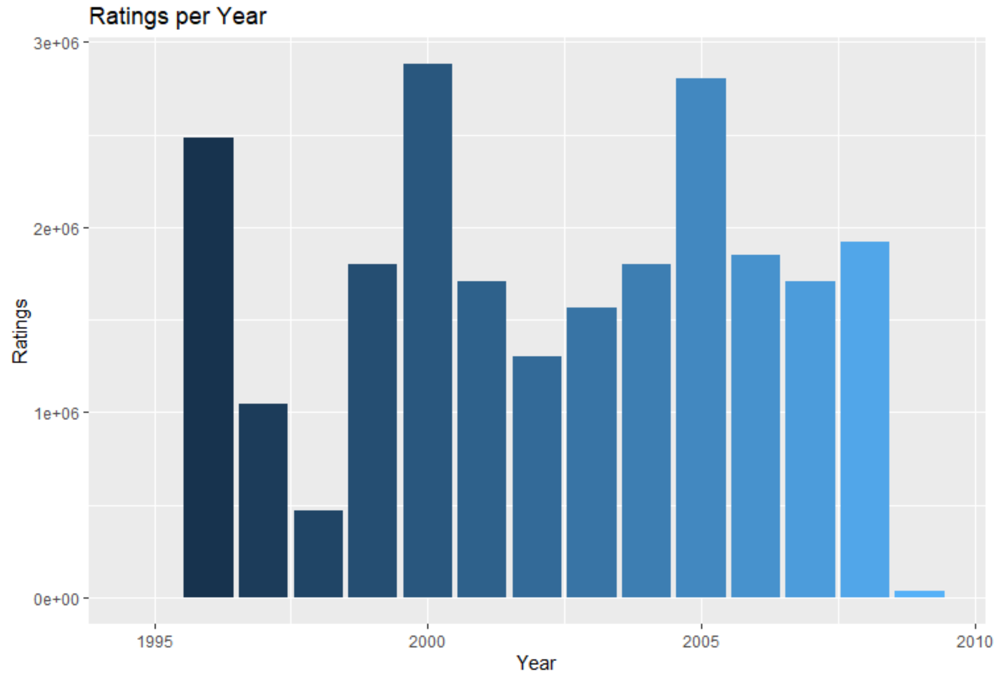
year	userId	ratings
2002	14463	9121
2007	67385	8834
2006	3817	6407
2007	47345	6181
2007	58357	5892
2002	7795	5701
2005	42791	5613
2006	30687	5479
2001	59269	5399
2006	31327	5343

Ratings

This plot shows us that the most of the ratings are above 3, and that the half-points are not so common,

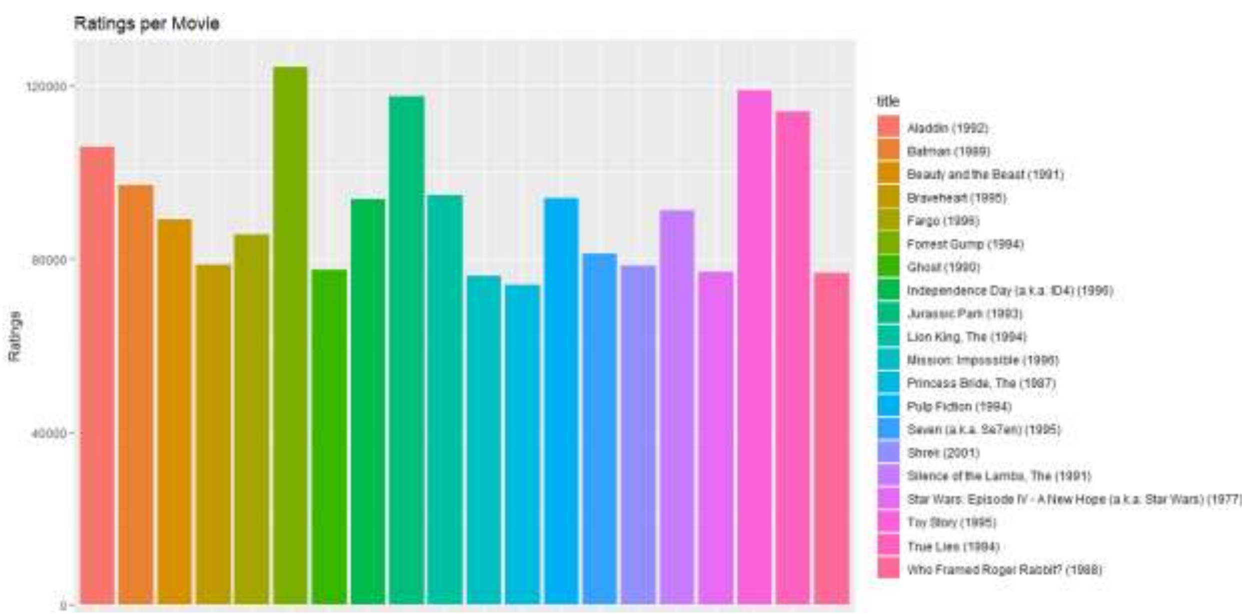


Ratings through years



Movies

Top 20 rated movies



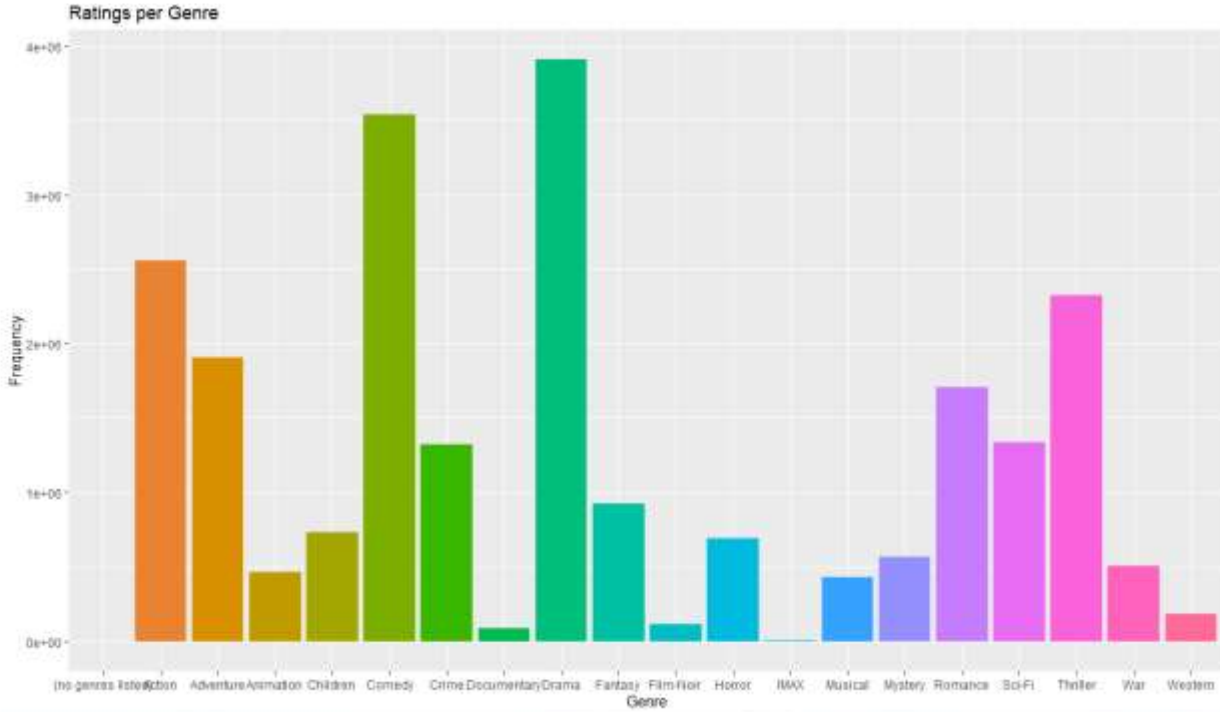
title	ratings
Forrest Gump (1994)	124316
Toy Story (1995)	118950
Jurassic Park (1993)	117440
True Lies (1994)	114115
Aladdin (1992)	105865
Batman (1989)	97108
Lion King, The (1994)	94605
Pulp Fiction (1994)	94086
Independence Day (a.k.a. ID4) (1996)	93796
Silence of the Lambs, The (1991)	91146
Beauty and the Beast (1991)	89145
Fargo (1996)	85580
Seven (a.k.a. Se7en) (1995)	81244
Braveheart (1995)	78636
Shrek (2001)	78378
Ghost (1990)	77440
Star Wars: Episode IV - A New Hope (a.k.a. Star Wars) (1977)	77016
Who Framed Roger Rabbit? (1988)	76825
Mission: Impossible (1996)	75968
Princess Bride, The (1987)	74045

These 58 movies were rated only once

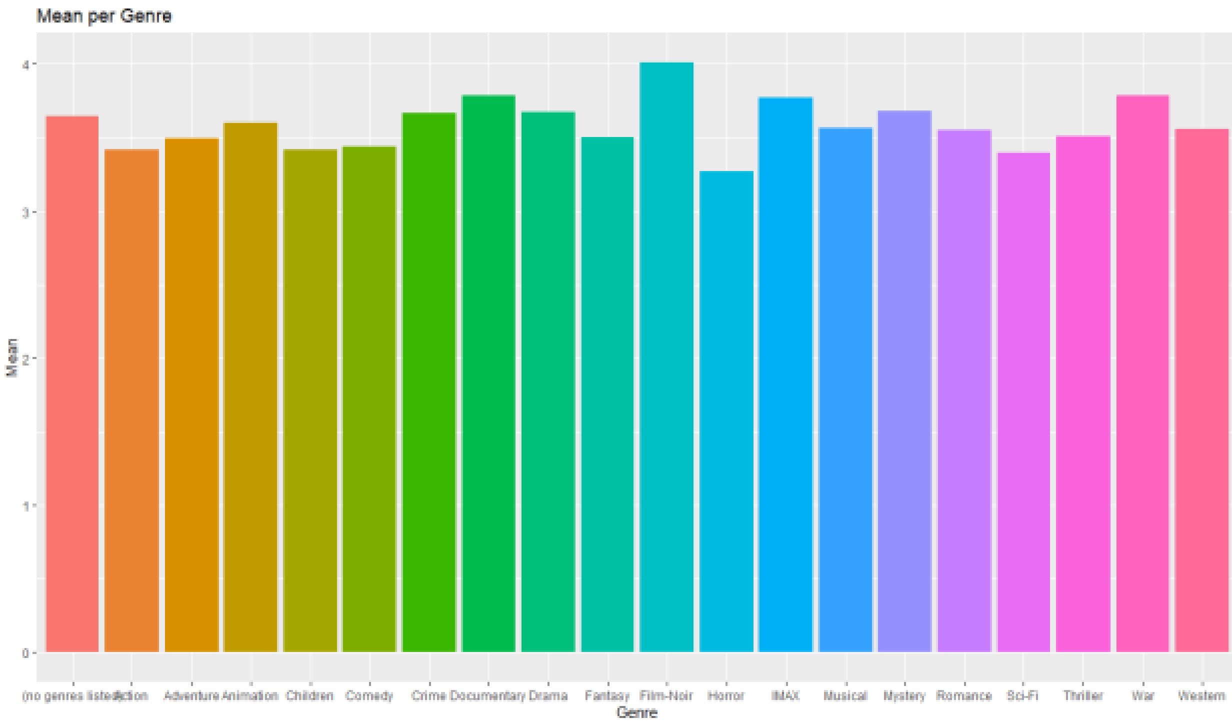
- 1, 2, 3, Sun (Un, deuz, trois, soleil) (1993)
- 4 (2005)
- Accused (Anklaget) (2005)
- Adios, Sabata (Indio Black, sai che ti dico: Sei un gran figlio di...) (1971)
- Africa addio (1966)
- Bellissima (1951)
- Blind Shaft (Mang jing) (2003)
- Brothers of the Head (2005)
- Condo Painting (2000)
- Confessions of a Superhero (2007)
- Cruel Story of Youth (Seishun zankoku monogatari) (1960)
- Deadly Companions, The (1961)
- Demon Lover Diary (1980)
- Devil's Chair, The (2006)
- Diminished Capacity (2008)
- Dog Run (1996)
- Dogwalker, The (2002)

- Du cÃ´tÃ© de la cÃ´te (1958)
- Face of a Fugitive (1959)
- Fireproof (2008)
- Fists in the Pocket (I Pugni in tasca) (1965)
- Flu Bird Horror (2008)
- Forgotten One, The (1990)
- Forty Shades of Blue (2005)
- God's Sandbox (Tahara) (2002)
- Guard Post, The (G.P. 506) (2008)
- Hellhounds on My Trail (1999)
- Hexed (1993)
- Hi-Line, The (1999)
- Hundred and One Nights, A (Cent et une nuits de Simon CinÃ©ma, Les) (1995)
- In the Winter Dark (1998)
- Jimmy Carter Man from Plains (2007)
- Just an American Boy (2003)
- Kanak Attack (2000)
- Ladrones (2007)
- Living 'til the End (2005)
- Love Forbidden (DÃ©fense d'aimer) (2002)
- Mala Noche (1985)
- Man Named Pearl, A (2006)
- Moonbase (1998)
- Mr. Wu (1927)
- Much Ado About Something (2001)
- Music Room, The (Jalsaghar) (1958)
- Part of the Weekend Never Dies (2008)
- Quarry, The (1998)
- Quiet City (2007)
- Relative Strangers (2006)
- SÃ¤g att du Ã¤lskar mig (2006)
- Stacy's Knights (1982)
- Stone Angel, The (2007)
- Symbiopsychotaxiplasm: Take One (1968)
- Testament of Orpheus, The (Testament d'OrphÃ©e) (1960)
- Tokyo! (2008)
- Train Ride to Hollywood (1978)
- Uncle Nino (2003)
- Variety Lights (Luci del varietÃ) (1950)
- Won't Anybody Listen? (2000)
- Young Unknowns, The (2000)

Genre



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


Mean per Genre
Mean
Genre
0
1
2
3
4
(no genres listed)
Action
Adventure
Animation
Children
Comedy
Crime
Documentary
Drama
Fantasy
Film-Noir
Horror
IMAX
Musical
Mystery
Romance
Sci-Fi
Thriller
War
Western

genres	mean
(no genres listed)	3.642857
Action	3.421405
Adventure	3.493544
Animation	3.600644
Children	3.418715
Comedy	3.436908
Crime	3.665925
Documentary	3.783487
Drama	3.673131
Fantasy	3.501946
Film-Noir	4.011625
Horror	3.269815
IMAX	3.767693
Musical	3.563305
Mystery	3.677001
Romance	3.553813
Sci-Fi	3.395743
Thriller	3.507676
War	3.780813
Western	3.555918

Predictive Model building and evaluation

First we are going to create a data frame to keep all the results in there

Naive Model

The first model and the simplest model is by predicting the mean that is:

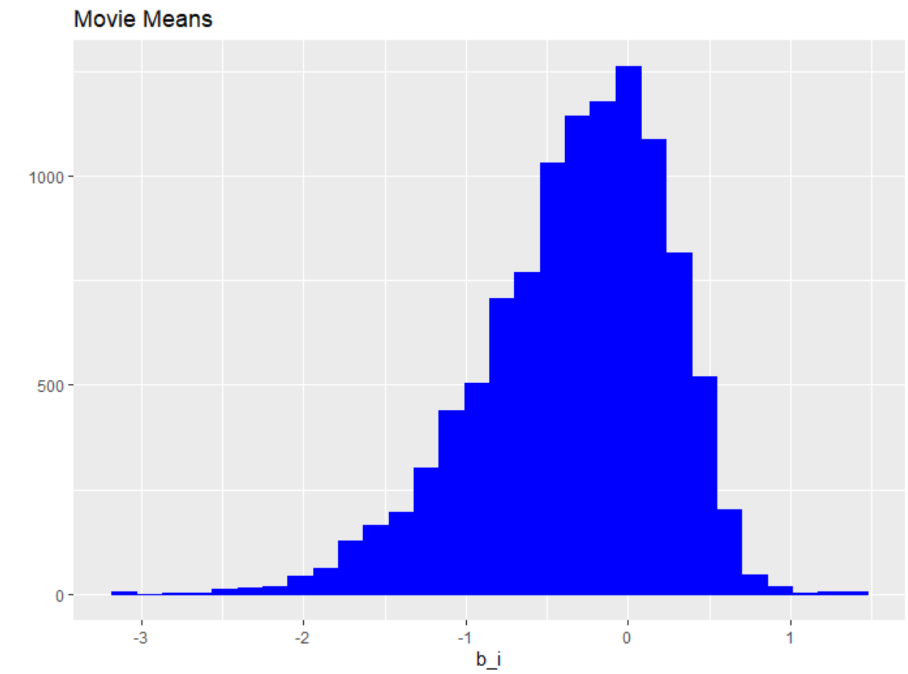
```
> mu_hat
[1] 3.527019
```

Using the validation dataset the RMSE is 1.052558, which bigger than our goal (below 0.86490)

model	variables	RMSE
Mean	All	1.052558

Movie based Model

Here we are going to take into account the movie

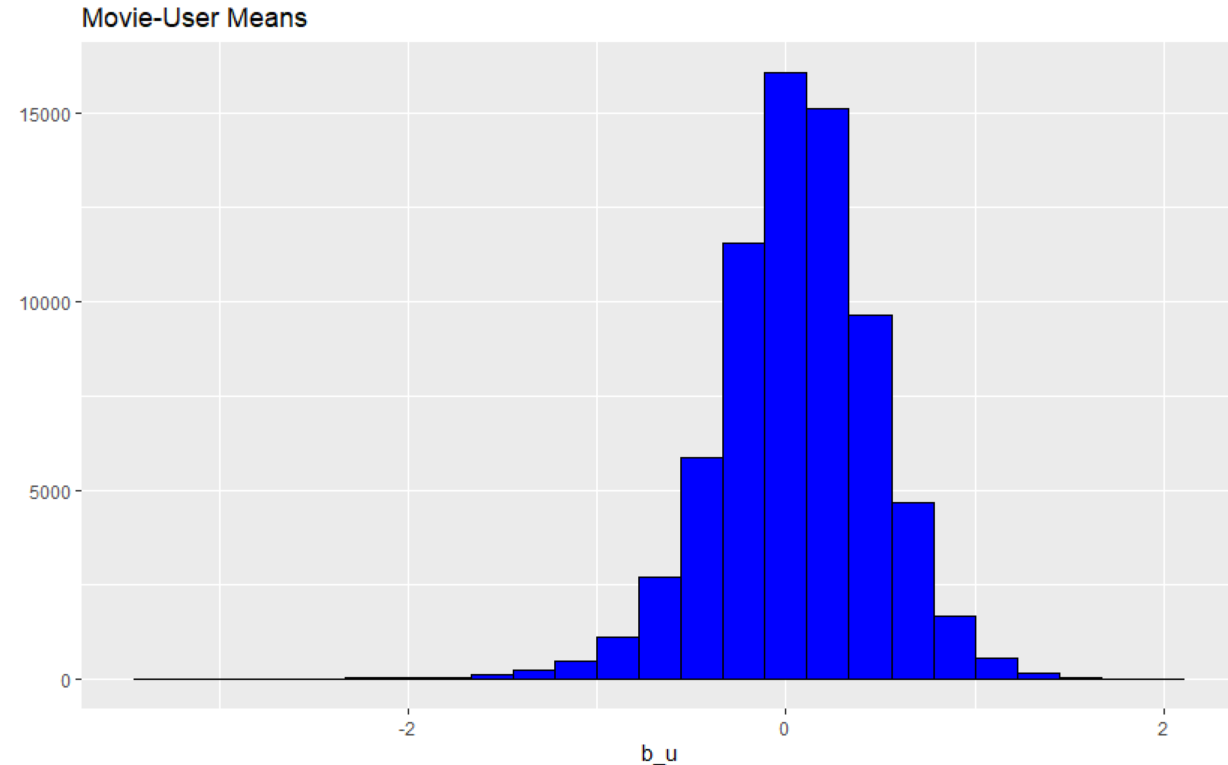


model	variables	RMSE
Mean	All	1.052558
Mean	Movie	0.941070

The RMSE on the validation dataset is 0.941070 which is also above our goal.

Movie and User Based Model

In this one we are going to take into account the user, in the exploratory phase we observed that each user rates different.

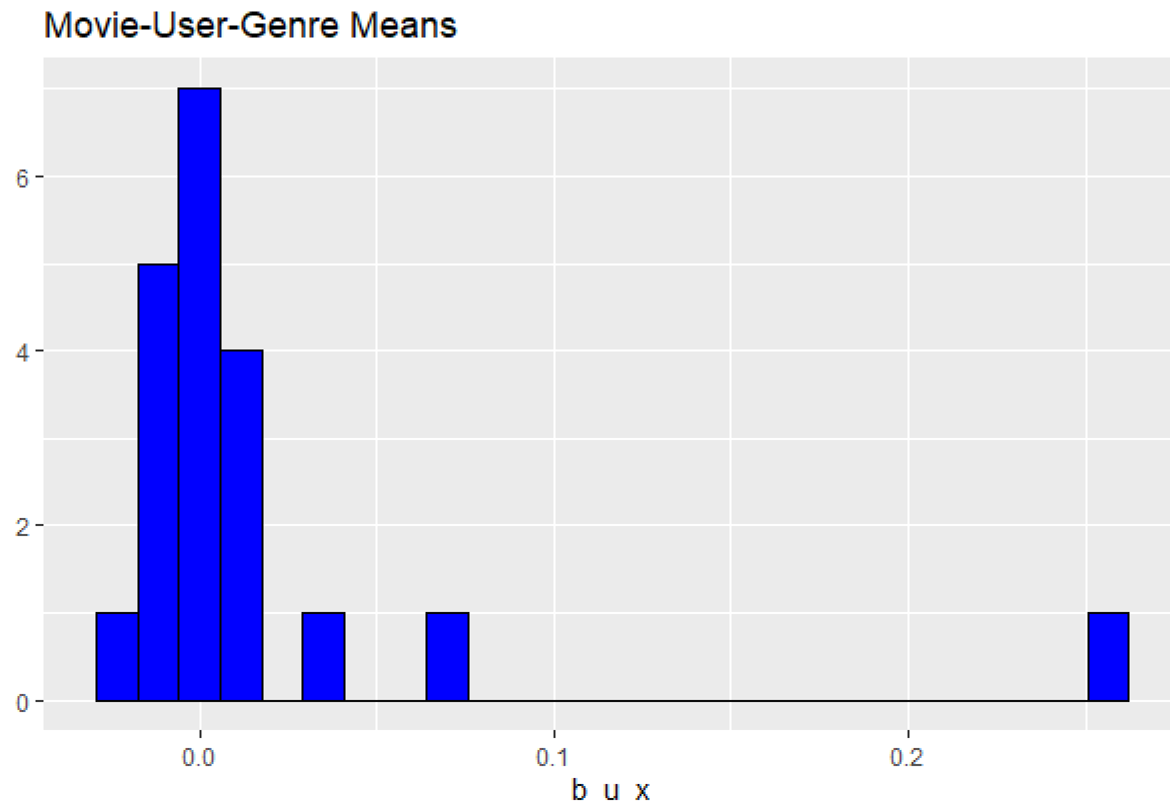


The result of the RMSE is 0.863366 which is already below our target!

model	variables	RMSE
Mean	All	1.052558
Mean	Movie	0.941070
Mean	Movie-User	0.863366

We are going to continue trying other models considering different variables and different combinations and see if we can improve our RMSE.

Movie, user and genre Based Model



model	variables	RMSE
Mean	All	1.0525579
Mean	Movie	0.9410700
Mean	Movie-User	0.8633660
Mean	Movie-User-Genre	0.8632723

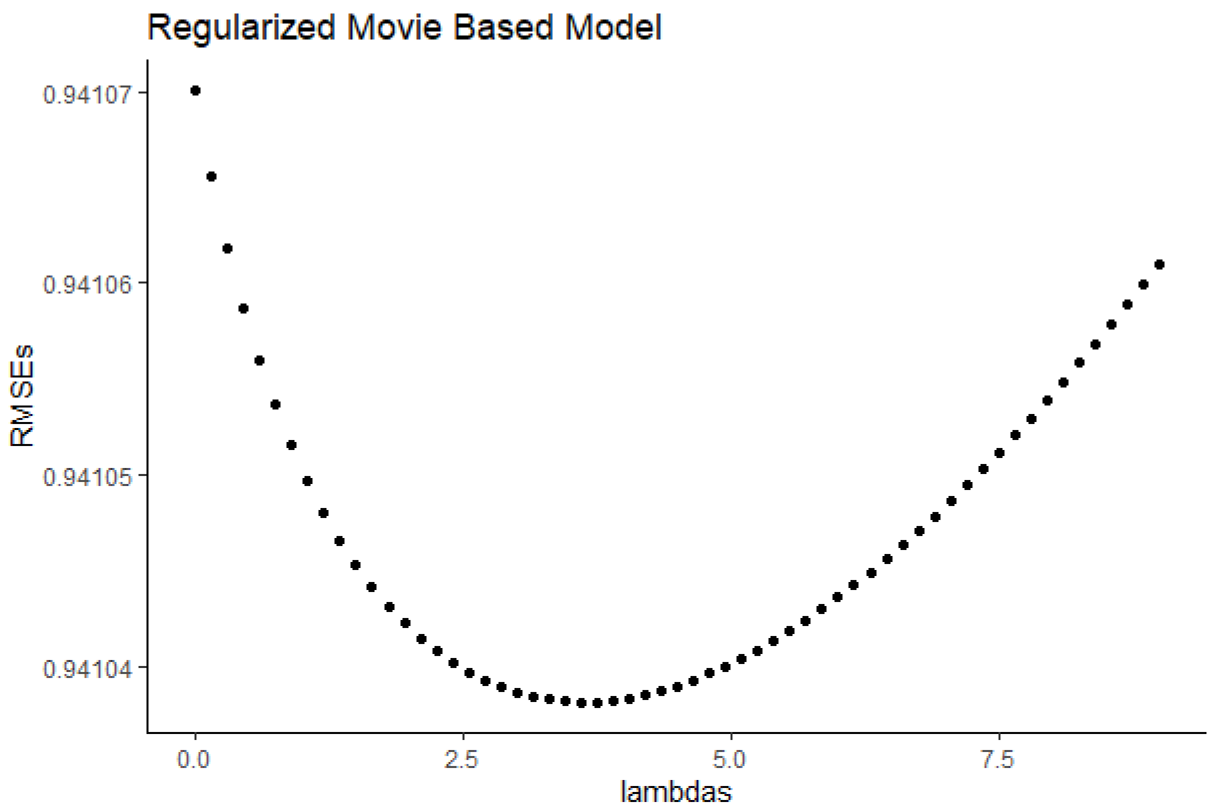
By adding the Genre, we get an even better RMSE, 0.8632723, so far is the best one.

Regularization

With regularization we penalize movies with large estimates that come from small sample sizes. First we are going to try a few values of lambda (penalty) and see which one works best.

lambdas <- seq(0, 9, 0.15)

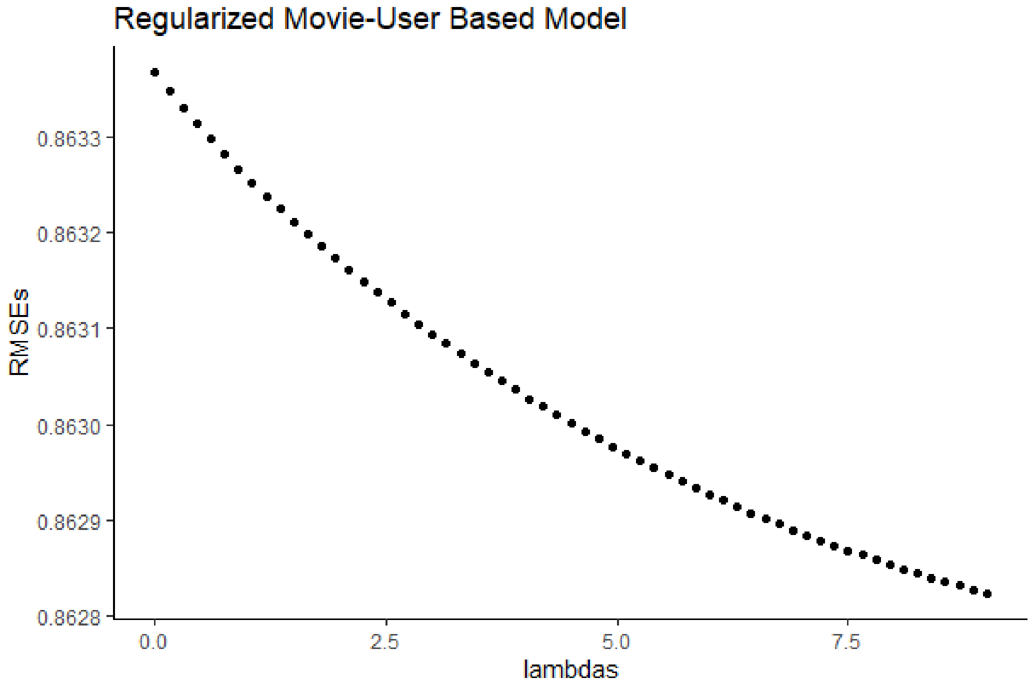
Movie



model	variables	RMSE
Mean	All	1.0525579
Mean	Movie	0.9410700
Mean	Movie-User	0.8633660
Mean	Movie-User-Genre	0.8632723
Regularized	Movie	0.9410381

The RMSE is 0.9410381 which is better than the one that wasn't regularized and only by taking into account the movie but is bigger than our target.

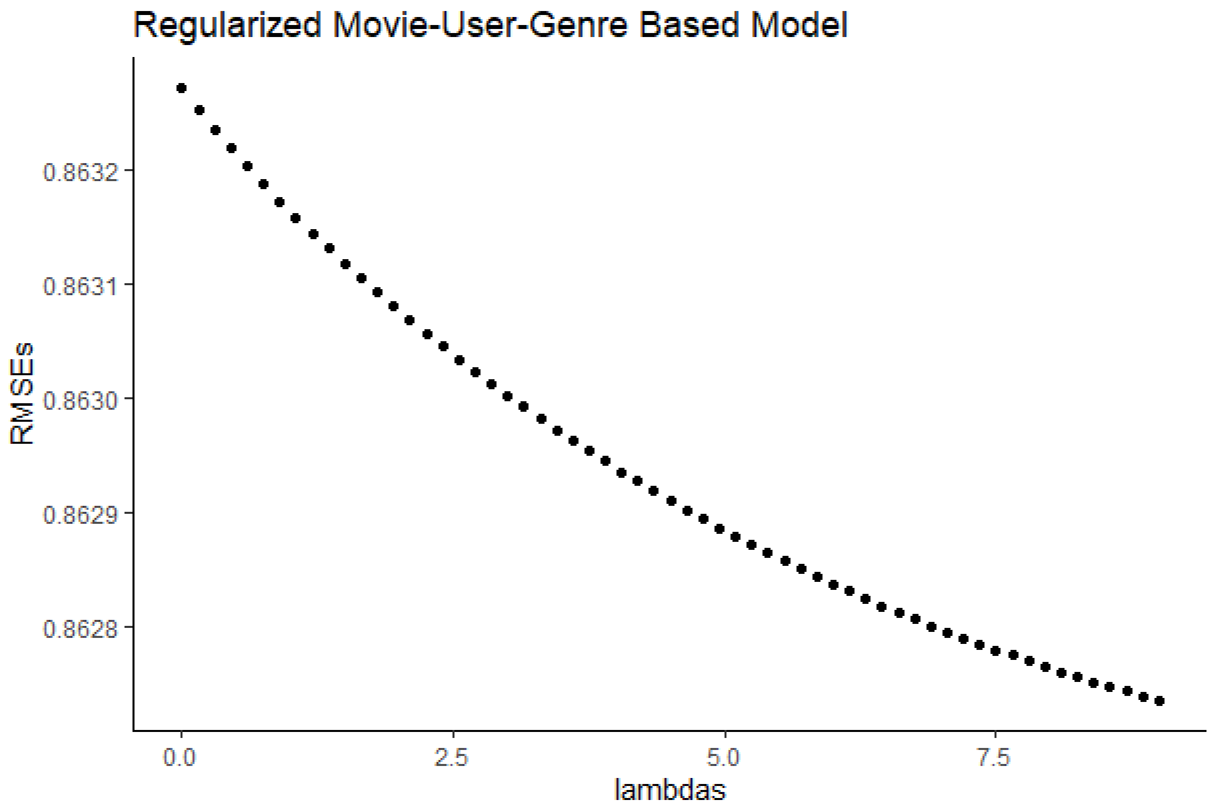
Movie and user



model	variables	RMSE
Mean	All	1.0525579
Mean	Movie	0.9410700
Mean	Movie-User	0.8633660
Mean	Movie-User-Genre	0.8632723
Regularized	Movie	0.9410381
Regularized	Movie-User	0.8628243

The RMSE on the validation dataset is 0.8628243 and this the best one gotten so far, it reaches our goal and is better than the Movie-User-Genre Based Line Model.

Movie, user and genre



model	variables	RMSE
Mean	All	1.0525579
Mean	Movie	0.9410700
Mean	Movie-User	0.8633660
Mean	Movie-User-Genre	0.8632723
Regularized	Movie	0.9410381
Regularized	Movie-User	0.8628243
Regularized	Movie-User-Genre	0.8627348

The RMSE on the validation dataset with this model is 0.8627348 which is the best one that we got with the built models. The genre does not improve much the results with the regularized and the non-regularized models but the Movie-User-Genre regularized model is the best one.

Results

This is the final table with the RMSE results of each model built.

model	variables	RMSE
Mean	All	1.0525579
Mean	Movie	0.9410700
Mean	Movie-User	0.8633660
Mean	Movie-User-Genre	0.8632723
Regularized	Movie	0.9410381
Regularized	Movie-User	0.8628243
Regularized	Movie-User-Genre	0.8627348

Conclusion

In this project first we prepare and wangle the data so that the type of the variables, the variables and everything is the way we wanted and easier to work with. Fortunately, the data did not need much wrangling. After a few steps it was ready to work with.

With the trained models and the variables we took into account we can conclude that the genre does not affect much the results like movie or user. But by taking it into account and regularized the model we get a RMSE of **0.8627348**, achieving our goal to get a RMSE lower than **0.86490**