Movielens Project

Contents

1 Introduction

In this project, I use Movielens data set to build a recommendation system based on movies' rating. All popular movie websites or streaming services like Netflix or Reddit apply recommendation systems to suggest viewers which movies they should choose. There are some methods to construct movie recommendation systems, for example content-based recommendation system, user-based collaborative filtering system, etc. In my project, the recommendation system is built based on movies' rating because we expect that users with similar taste will tend to rate movies with high correlation. Therefore, the main objective is construct a model to predict user movie ratings based on other users' ratings.

The structure of the report is as follows: Section 1 introduces the analytic problem, Section 2 presents data preparation, data exploration and visualization are included in Section 3, Section 4 discusses methodologies and results, finally a conclusion with limitations and further analysis is presented in Section 5.

2 Data preparation

2.1 Data description

MovieLens data sets are used to build rating prediction model in this project, they were collected by the GroupLens Research Project at the University of Minnesota over various periods of time, depending on the size of the set. I use the 10M version of the MovieLens data set for this project and create a data set with 7 variables, including userId, movieId, rating, timestamp, title, genres and year (I separate "year" variable to represent for "released year of the movie" from "genres" variable). I start by explore the overall of the data set. The summary of this data set is presented in Table 1 and Table 2.

Table 1: Summary of character variables

Character_variables	Number_of_categories
title	10407
genres	797

Table 2: Summary of numeric variables

Numeric_variables	Min	Median	Mean	Max	NA_number
userId	1.0	35738	35869.8	71567	0
movieId	1.0	1834	4121.7	65133	0
rating	0.5	4	3.5	5	0
timestamp	789652009.0	1035493918	1032615907.4	1231131736	0
year	1915.0	1994	1990.2	2008	0

From the summary of the data set, we see that there are 7 variables, including 2 character and 5 numeric ones. Because the objective is building a model to predict rating, the response variable here is "rating" and predictors are remaining variables. Specifically, the response variable "rating" has a range from 0.5 to 5, with mean value is around 3.5. Moreover, there is no N/A value in the data set, thus I do not need to deal with missing data.

2.2 Data cleaning

In this section, we investigate each variable and check if there is any outlier in the data set. Because 5 variables, including userId, movieId, timestamp, title and genres are actually factor values, I explore the boxplots of rating and year only in Figure 1.

Figure 1 shows that it seems to be no outlier in rating and year variables, just only some points are far from the center but still in the value range of each variable.

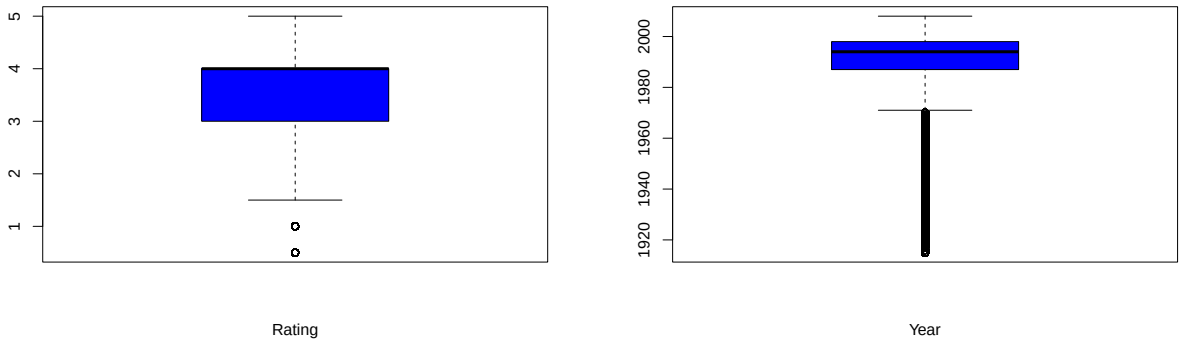

Figure 1: Boxplots of rating and year

3 Data exploration and visualization

3.1 Rating count exploration

In the upcoming sections, I explore more details in the data set and also visualize if possible. There are 10,677 different movies in this data set, with the rating count ranges from 1 to 65,133, which means that there is a significant difference in rating count among movies. To understand more about the difference in rating count, I visualize by plotting the histogram of number of rating count in Figure 2.

Please note that because the tail of histogram with movies having more than 3000 rating counts will be hardly to see, I just plot histogram for movies having less than 3000 rating counts only.

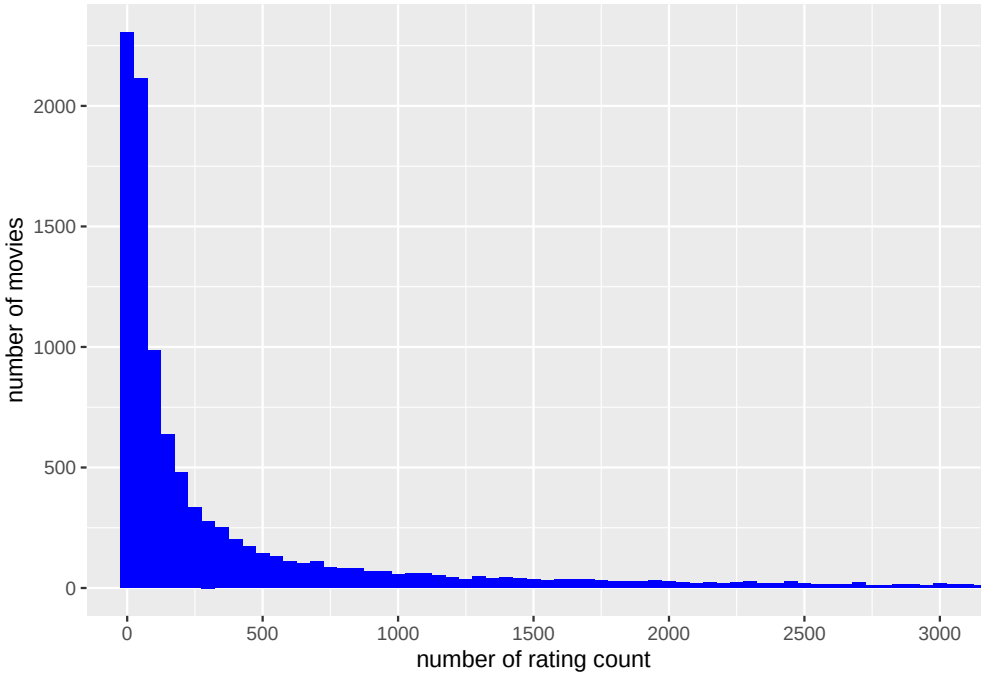

Figure 2: Histogram of rating count

The distribution of rating count is right skewed because the right tail (larger values) of histogram is much longer than the left tail (smaller values), which means that most values have small number of rating count.

Next I discover the top 10 movies with highest number of rating count. The top 10 list is shown in Figure 3.

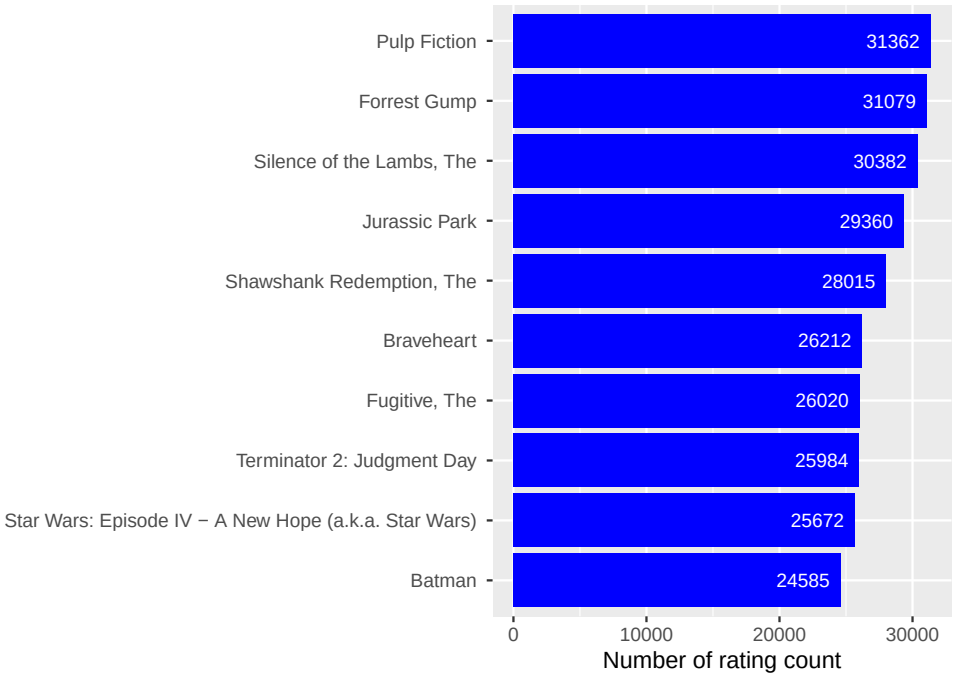

Figure 3: Top 10 movies with highest rating count

Figure 3 makes sense since we can see most familiar movie names here. However, there is a big difference in number of rating count among the top 10 highest rating movies.

3.2 Average rating exploration

Analysis of average rating is presented in this section. To begin with, I visualize the average rating by plotting a histogram in Figure 4.

We can see from Figure 4 that the distribution of average rating is left skewed, with the mode value around 3.5. However, not many movies reach the highest rating at 5. In Table 3, we discover the top 10 movies with highest average rating. We can see from Table 3 that although these movies rated with highest points, actually they are not popular movies with few viewers only. Hence, they are not representative for the data set.

Table 3: The top 10 movies with highest average rating

title	rating_count	average_rating
Blue Light, The (Das Blaue Licht)	1	5.00
Fighting Elegy (Kenka erejii)	1	5.00
Hellhounds on My Trail	1	5.00
Satan's Tango (SÃ¡tÃ¡ntangÃ³)	2	5.00
Shadows of Forgotten Ancestors	1	5.00
Sun Alley (Sonnenallee)	1	5.00
Constantine's Sword	2	4.75
Human Condition II, The (Ningen no joken II)	4	4.75
Human Condition III, The (Ningen no joken III)	4	4.75
Who's Singin' Over There? (a.k.a. Who Sings Over There) (Ko to tamo peva)	4	4.75

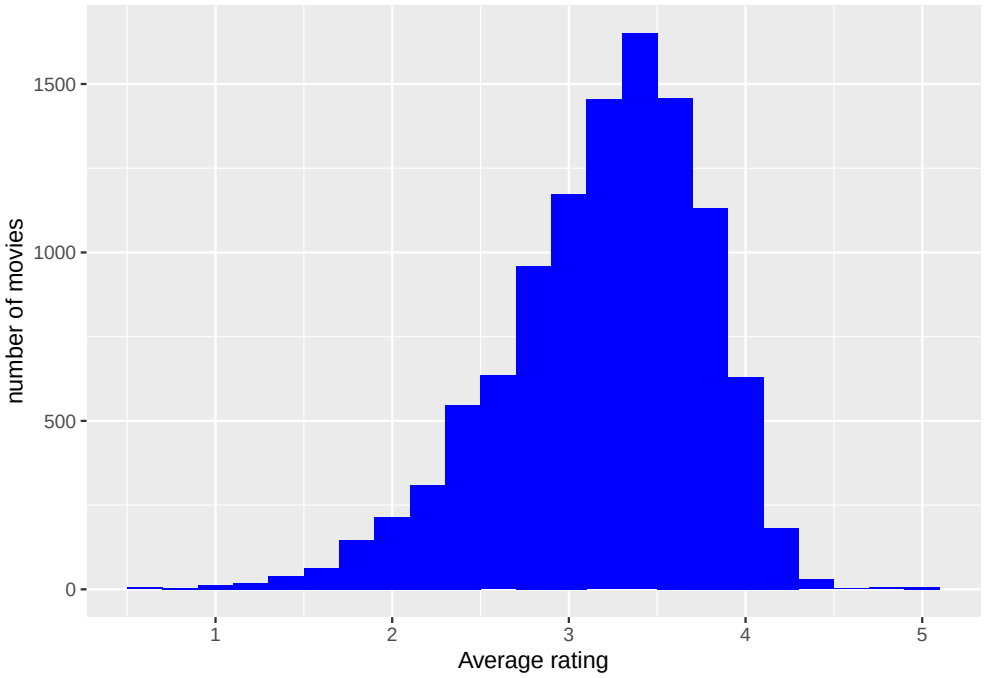

Figure 4: Histogram of average rating

3.3 Genres exploration

In this section, I explore movies by genres. Because there is a combination of different genres in the genres variable, so I separate this variable into each genre and discover it. After breaking down genres variable into separate genres, there are 20 different genre names. The number of movies in each genre are presented in Figure 5.

It can be seen from Figure 5 that drama movies are the most popular type, followed by comedy and action types. On the other hand, much less viewer choose IMAX and documentary movies than other types.

However in the data set, each movie is usually is a combination of different genres,thus I explore more details of rating count and average rating by the combination of genres. Although there are only 20 separate genre, it can generate 797 combination of genres in the data set. Table 4 shows the highest rating count and we can see that drama and comedy types have the higher number of rating count than other kinds of movies.

Table 4: The top 10 highest rating count by combination of genres

genres	genres_count
Drama	733296
Comedy	700889
Comedy|Romance	365468
Comedy|Drama	323637
Comedy|Drama|Romance	261425
Drama|Romance	259355
Action|Adventure|Sci-Fi	219938
Action|Adventure|Thriller	149091
Drama|Thriller	145373
Crime|Drama	137387

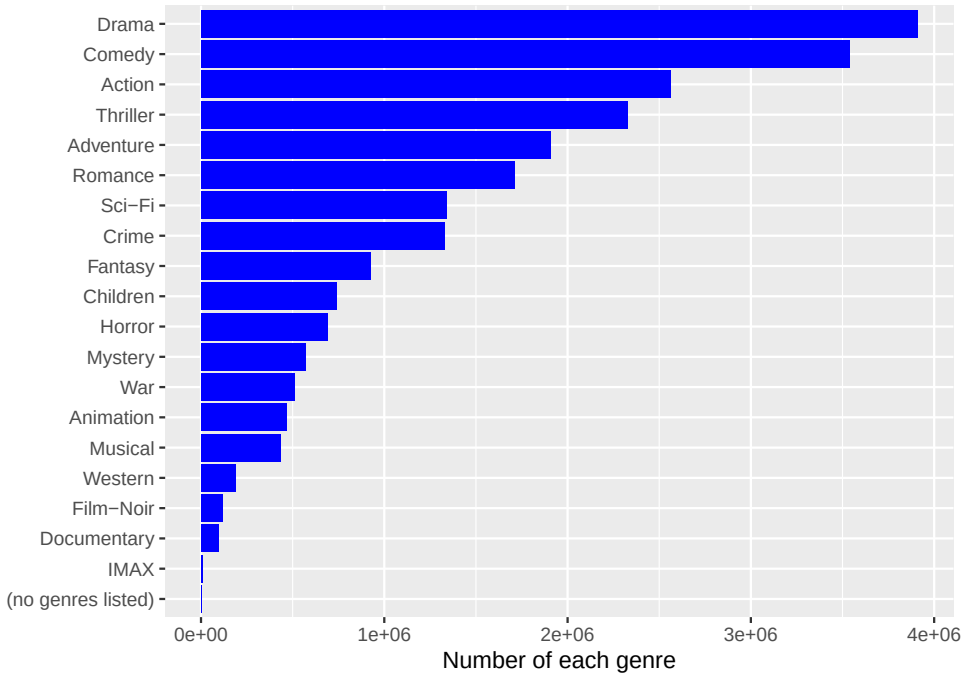

Figure 5: Total number of movie by genre

Figure 6 presents the top 10 highest average rating by combination of genres. Although drama movies has the highest number of rating count, they do not have the highest average rating. Movies with combination of Animation|IMAX|Sci-Fi genres have highest average rating at 4.7 point.

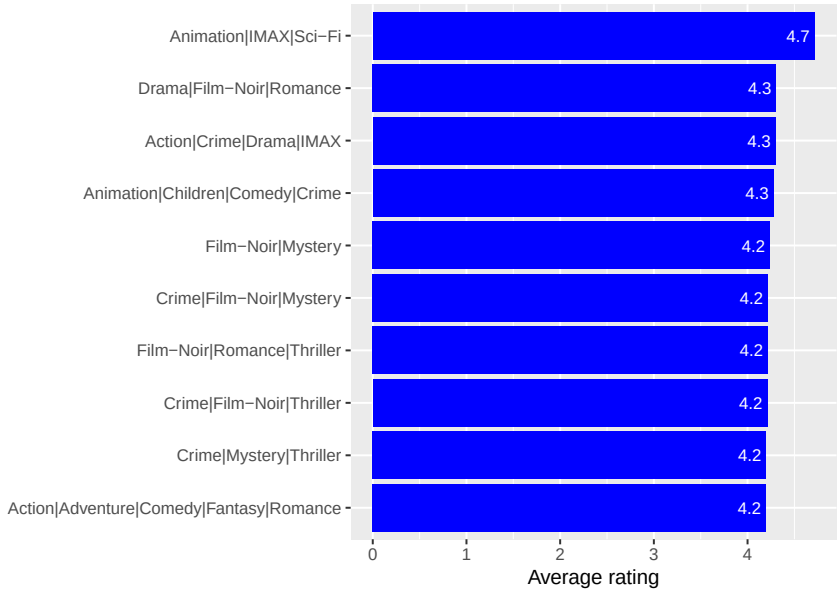

Figure 6: Top highest average rating by combination of genres

3.4 Released year exploration

As mentioned in the Introduction part, I separated released year from genres variable into "year" variable. In this section, I go into deep understanding of this data set by year. Figure 7 shows the number of movies released in each year.

It can be seen from Figure 7 that the number of movies increases gradually from 1915 to 1990, but soars significantly from 1990 to 1995 and peaks at 1995. After that, the number drops from 1995 to 2008.

Table 5: The top 10 highest number of movies by year

year	count	average_rating
1995	786762	3.442880
1994	671376	3.472206
1996	593518	3.361044
1999	489537	3.453140
1993	481184	3.496722
1997	429751	3.363998
1998	402187	3.415834
2000	382763	3.395192
2001	305705	3.439049
2002	272180	3.450694

Table 5 shows more details about the top 10 years with highest number of movies, their average ratings are around 3.4. Both Figure 7 and Table 5 indicate that the highest number of movies released in 1995, followed by 1994 and 1996.

In the last section in Part 3, I explore the top 10 years with highest average rating. It is a little bit surprise here as we can see from Table 6 that the years with the highest average rating are quite long time ago and there are not many viewers in those years.

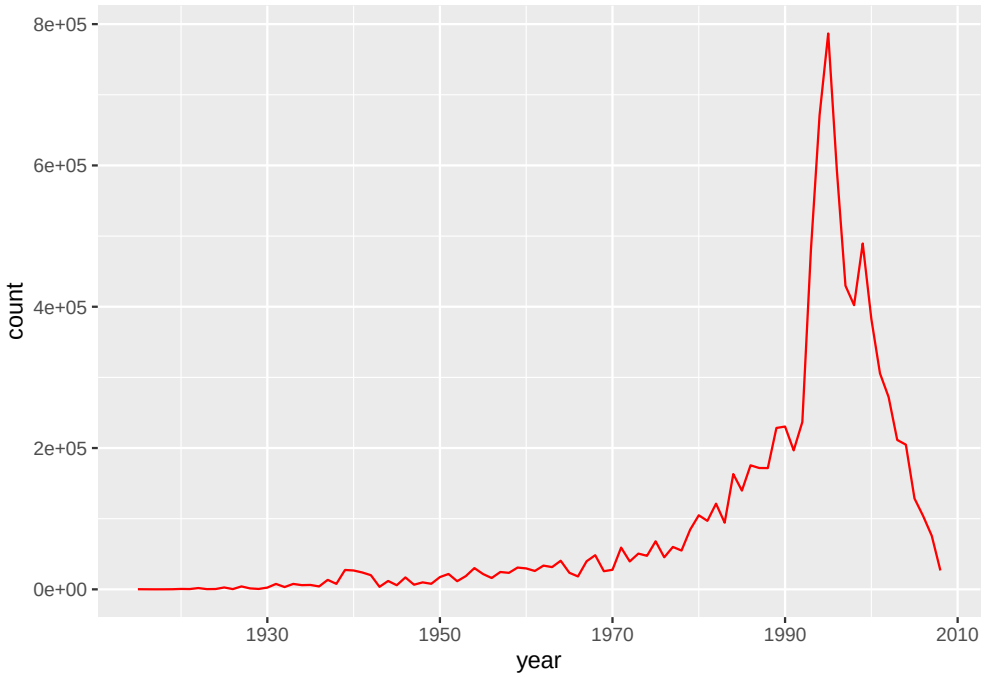

Figure 7: Number of movies each year

Table 6: The top 10 years with highest average rating

year	average_rating	view_count
1946	4.053341	16882
1934	4.050974	5954
1942	4.044985	20051
1931	4.025232	7669
1941	4.018256	23883
1927	4.015364	4133
1954	4.008591	30089
1957	4.008572	24557
1944	3.990309	11918
1962	3.989813	33622

4 Methodology

4.1 Method to predict rating

In this analysis, I build a regression model to predict rating based on predictors in the data set. Moreover, as we can see from Section 3, there are some highest or lowest points in rating data but not representative for all data set, thus I also apply regularization method to produce better prediction. Some different models are considered and the optimal one is chosen based on the result of Root Mean Squared Error (RMSE). The equation of RMSE is as follows:

$$RMSE = \sqrt{1/N * \sum_{u,i} (\hat{y}_{u,i} - y_{u,i})^2} \tag{1}$$

where $y_{u,i}$ is the rating for movie i by user u, $\hat{y}_{u,i}$ is the predicted value of $y_{u,i}$ and N is the number of user/movie combinations and the sum occurring over all these combinations.

Firstly, I estimate rating by the model with no predictor, which means the same rating for all movies. The function is like that:

$$\hat{y}_{u,i} = \mu + \epsilon_{u,i} \tag{2}$$

with $\epsilon_{u,i}$ is independent errors sampled from the same distribution centered at 0 and μ is the mean rating of all movies.

Model 1

As we can see from Section 3 that there is a big difference of rating across genres. In other words, movies with different genres are rated differently. Hence, I add the term b_g to represent for genres effect into Equation (2).

$$\hat{y}_{u,i} = \mu + b_g + \epsilon_{u,i} \tag{3}$$

The idea of regularization is to constrain the total variability of the effect sizes, so instead of minimizing RMSE in Equation (1), we minimize an equation that adds a penalty in Equation (4).

$$1/N \sum_{u,i} (y_{u,i} - \mu - b_g)^2 + \lambda \sum_{g} b_g^2 \tag{4}$$

It can be seen as penalized least squares, including least squares and penalty that gets larger when many b_g are large. The values of b_g that minimize Equation (4) are shown in Equation (5).

$$\hat{b_g}(\lambda) = 1/(\lambda + n_g) \sum_{i=1}^{n_g} (Y_{u,i} - \hat{\mu}) \tag{5}$$

where n_g is the number of ratings made for genre g.

Model 2

Section 3 also shows that rating of movies different across years. Therefore, to account for year effect, I add the term b_y to represent for year effect into Equation (3).

$$\hat{y}_{u,i} = \mu + b_g + b_y + \epsilon_{u,i} \tag{6}$$

The equation of penalized least squares is:

$$1/N \sum_{u,i} (y_{u,i} - \mu - b_g - b_y)^2 + \lambda (\sum_{g} b_g^2 + \sum_{y} b_y^2) \tag{7}$$

The values of b_y that minimize Equation (7) can be approximated by:

$$\hat{b_y}(\lambda) = 1/(\lambda + n_y) \sum_{i=1}^{n_y} (Y_{u,i} - \hat{\mu} - \hat{b_g}(\lambda)) \tag{8}$$

where n_y is the number of ratings made for year y.

Model 3

The rating of movies is also affected by the movie itself, thus I add movie effect inside model, the equation is as below:

$$\hat{y}_{u,i} = \mu + b_g + b_y + b_i + \epsilon_{u,i} \tag{9}$$

The equation of penalized least squares is:

$$1/N \sum_{u,i} (y_{u,i} - \mu - b_g - b_y - b_i)^2 + \lambda (\sum_g b_g^2 + \sum_y b_y^2 + \sum_i b_i^2) \tag{10}$$

The values of b_i that minimize Equation (10) can be approximated by:

$$\hat{b}_i(\lambda) = 1/(\lambda + n_i) \sum_{i=1}^{n_i} (Y_{u,i} - \hat{\mu} - \hat{b_g}(\lambda) - \hat{b_y}(\lambda)) \tag{11}$$

where n_i is the number of ratings made for movie i.

Model 4

Finally, different viewers usually rate the same movies with different point. Hence, I include user effect in prediction model, the equation is as below:

$$\hat{y}_{u,i} = \mu + b_g + b_y + b_i + b_u + \epsilon_{u,i} \tag{12}$$

The equation of penalized least squares is:

$$1/N \sum_{u,i} (y_{u,i} - \mu - b_g - b_y - b_i - b_u)^2 + \lambda (\sum_g b_g^2 + \sum_y b_y^2 + \sum_i b_i^2 + \sum_u b_u^2) \tag{13}$$

The values of b_u that minimize Equation (13) can be approximated by:

$$\hat{b}_u(\lambda) = 1/(\lambda + n_u) \sum_{i=1}^{n_u} (Y_{u,i} - \hat{\mu} - \hat{b_g}(\lambda) - \hat{b_y}(\lambda) - \hat{b}_i(\lambda)) \tag{14}$$

where n_u is the number of ratings made by user u.

I split edx data set into two set for training and testing (with 90% and 10% of edx set respectively) then use train set to train each model and test set to calculate RMSE of each model. Furthermore, because λ is a tuning parameter, I use cross validation to choose the optimal value of *lambda* with its range from 0 to 10. Based on that, I choose the optimal model with smallest RMSE and use this chosen model to retrain on the whole edx set to find the final RMSE on validation set.

4.2 Selecting the optimal model

As mentioned in the previous part, I run regress model with regularization method on on the training set and use the test set to compare RMSE results, then choose the optimal model based on that. The final result is presented in Table 7.

Table 7 shows that the model with only genre effect produces quite high RMSE and RMSE reduces just a little bit by adding year effect into model. It then drops significantly when including 2 more movieId and userId variables, gives much better results. Therefore, I choose the optimal model including all 4 effects: genres, year, movieId and userId.

Table 7: Combined result of all models

Method	Optimal_lambda	RMSE
Genre effect model	2	1.0180938
Genre + year effect model	4	1.0105835
Genre + year + movieId effect model	2	0.9433988
Genre + year + movieId + userId effect model	5	0.8644398

Figure 8 presents different values of lambda and respectively RMSE of Model 4 with all 4 effects. Because Model 4 is chosen, I will set value of lambda equal 5 as the optimal lambda to calculate RMSE on the validation set.

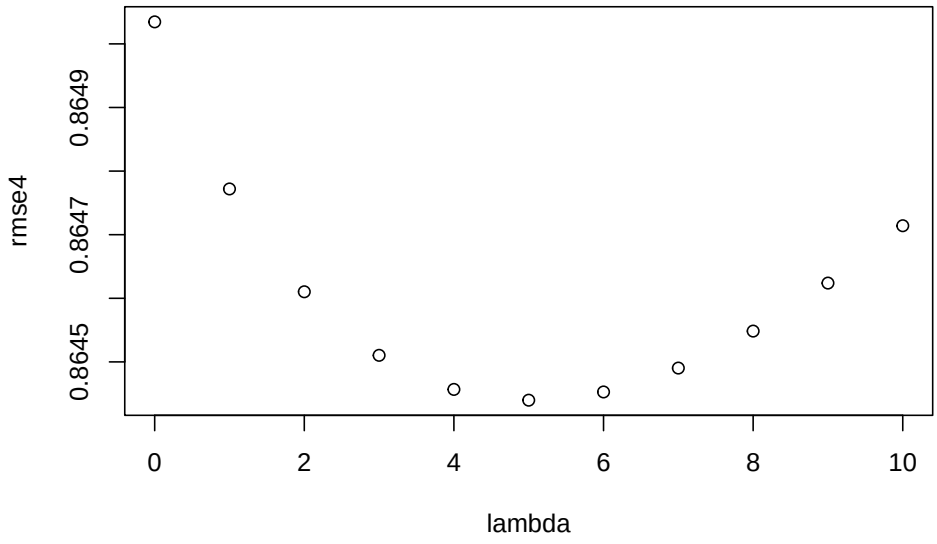

Figure 8: Value of lambda and RMSE of Model 4

4.3 Final result on validation set

I retrain the chosen model with value of optimal lambda equal 5 on whole edx set, then use the validation set to get the final quality of model (RMSE) in this section. The result is shown in Table 8.

Table 8: Final RMSE result on validation set

Method	RMSE
Genre + year + movieId + userId effect model	0.8647846

The final RMSE result on the validation set is not much different from the result on the test set and much less than 1 (i.e., the standard error of rating), which we can suppose that it gives a good prediction for rating of a certain movie. Hence, based on that, the movie recommendation system works effectively.

5 Conclusion

Analysing the movieLens data set gave many interesting insights into the movie business. By exploration and visualization, we can find out more trends in the movie data set. In this project, I just use a simple linear regression to predict rating and only few variables are included into model. I think that we can extend to more complicated models with more variables which are available in the full data set, which will help to produce a better prediction for a movie rating.