MovieLens

George Seymour Denis

6/22/2020

Overview

This file reports the movielens project. The goal is to predict movie ratings using the parameters provided in the movielens data set.

Important Notice: The title and genres columns of the edx and validation data sets downloaded from the grouplens.org website have no useful information (NA's) after loading them with the provided script. Therefore, the edx and validation data sets used here are provided by edx staff member @wonyoungcheong via the following link:

https://drive.google.com/drive/folders/1IZcBBX0OmL9wu9AdzMBFUG8GoPbGQ38D?usp=sharing

There are 2 .rds files within the above link, one has the edx data set, the other has the validation data set, both recognizable within the file names They are downloaded directly into the computer, specifically in the current working directory to make them easier to load. The R version used for the project is 4.0 !! Some training process could not be performed due to low computer memory space !!

Analysis

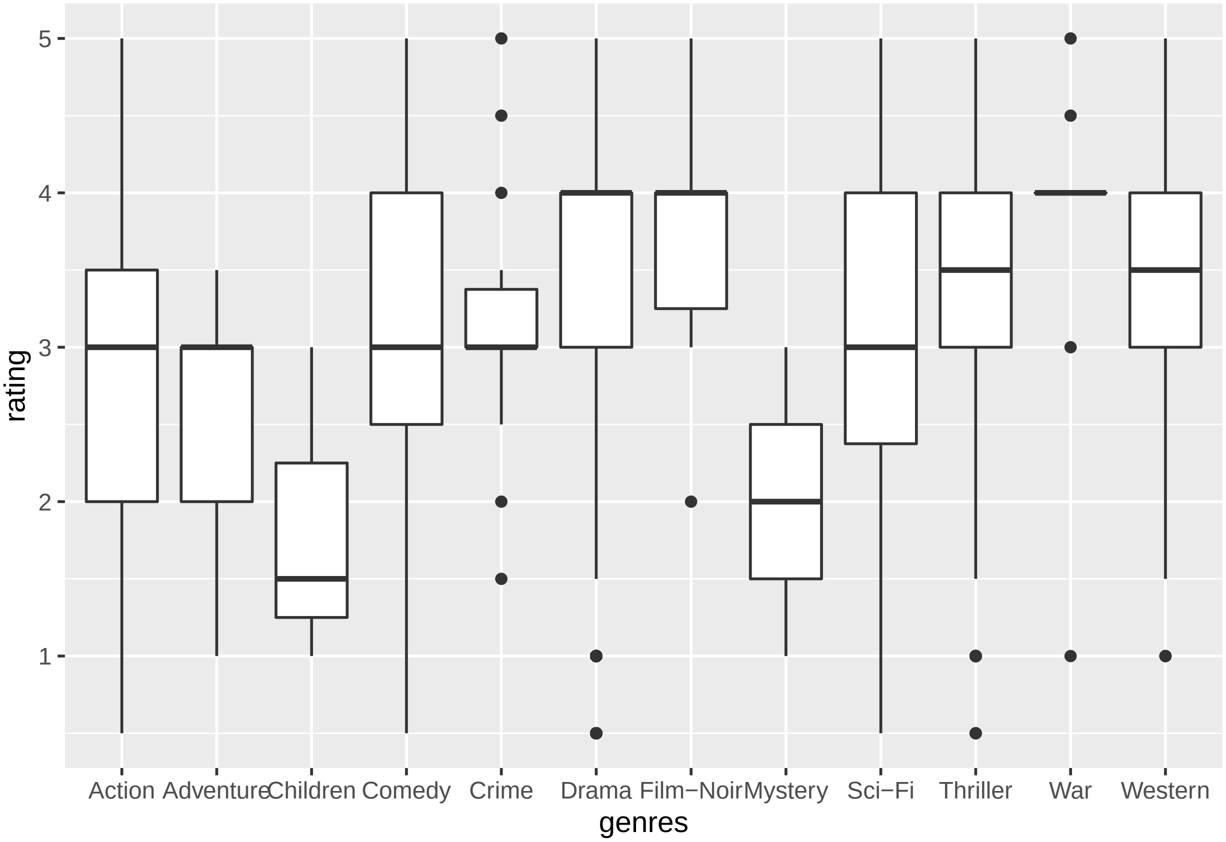
Movie rating levels are basically the effects of the user's point of view and on how the movie performed. The viewer may basically like action, thriller, drama and sci-Fi movies for example, so he initially thinks of giving a five (5) star rating for these type of movies, but it happens that some of them don't meet it's expectations, in this case, he intends to decrease that rating depending on how much he has been satisfied.This is how most people rate not only movies, but also other things like articles they have purchased online for example. There are some people who don't like very long movies, and some who don't like the short ones, but there are a lot movie fans for whom the running time of the movie doesn't matter. In this case, taking the movielens project into account, we are going to analyse our movie rating within the userId, the title, and the genres variables.

We load the edx data set, set the seed to 1 and devide it into two parts with 90% for the test set, then we are going to see average and the total ratings for the userId, the title and the genres one by one. The results are saved in edx_train_ura, edx_train_mra and edx_train_gra respectively.

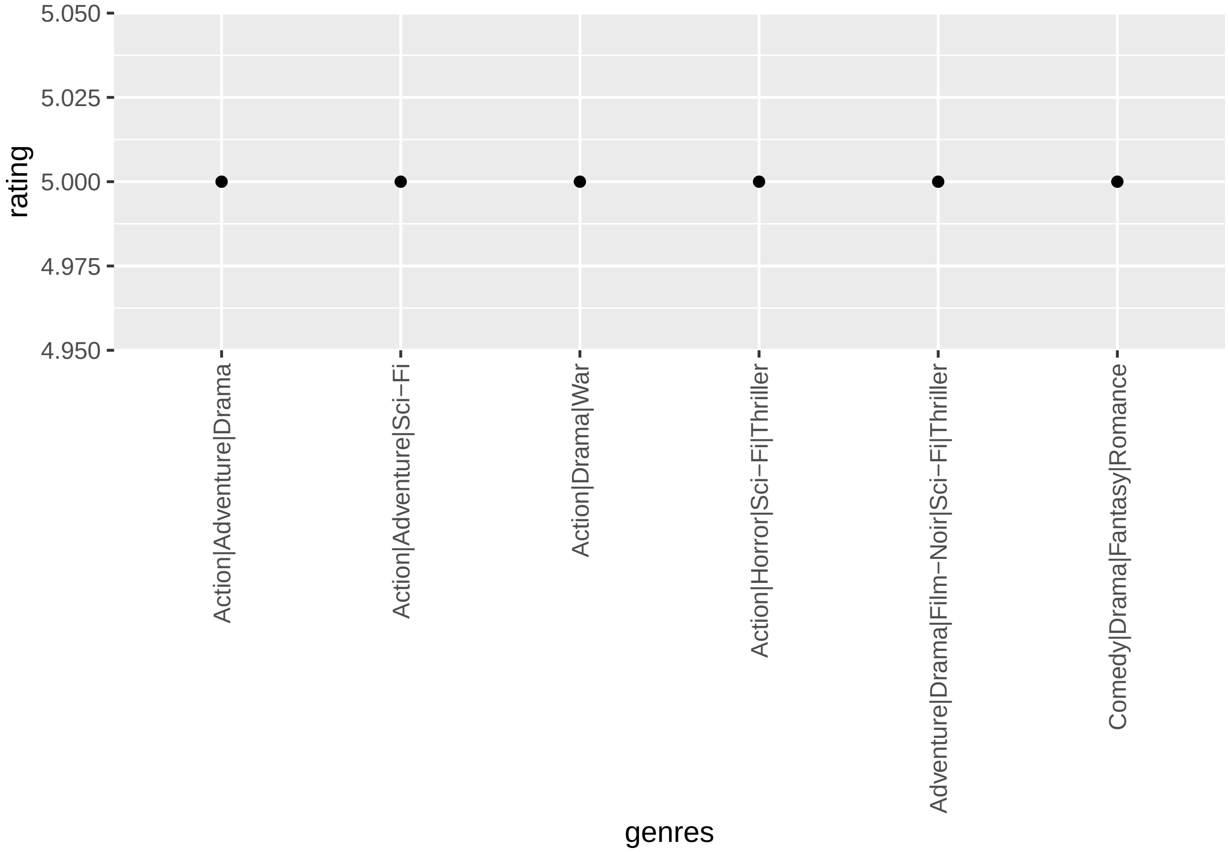
To avoid repeating the same userIds genres and titles, we use the unique function so the resulting data set will be more readable. We then observe each of them by viewing each of them.

We first view the highest ratings in genres. We notice that the Action, Comedy, Drama genres are at the top of the list with average rating surrounding 4.6. The Comedy and Drama genres have the highest number of ratings. Another thing is, when we are looking at the single Drama genres, we see that it has an average rating of 3.71, the single Comedy genres has an average of 3.24, the single Comedy genres has an average rating of 3.24, but the combination of these two reaches a higher level of 3.61. In other words, the genres of the movie have some effects but don't influence the rating very much. Movies may have the same genres, but every movie is unique.

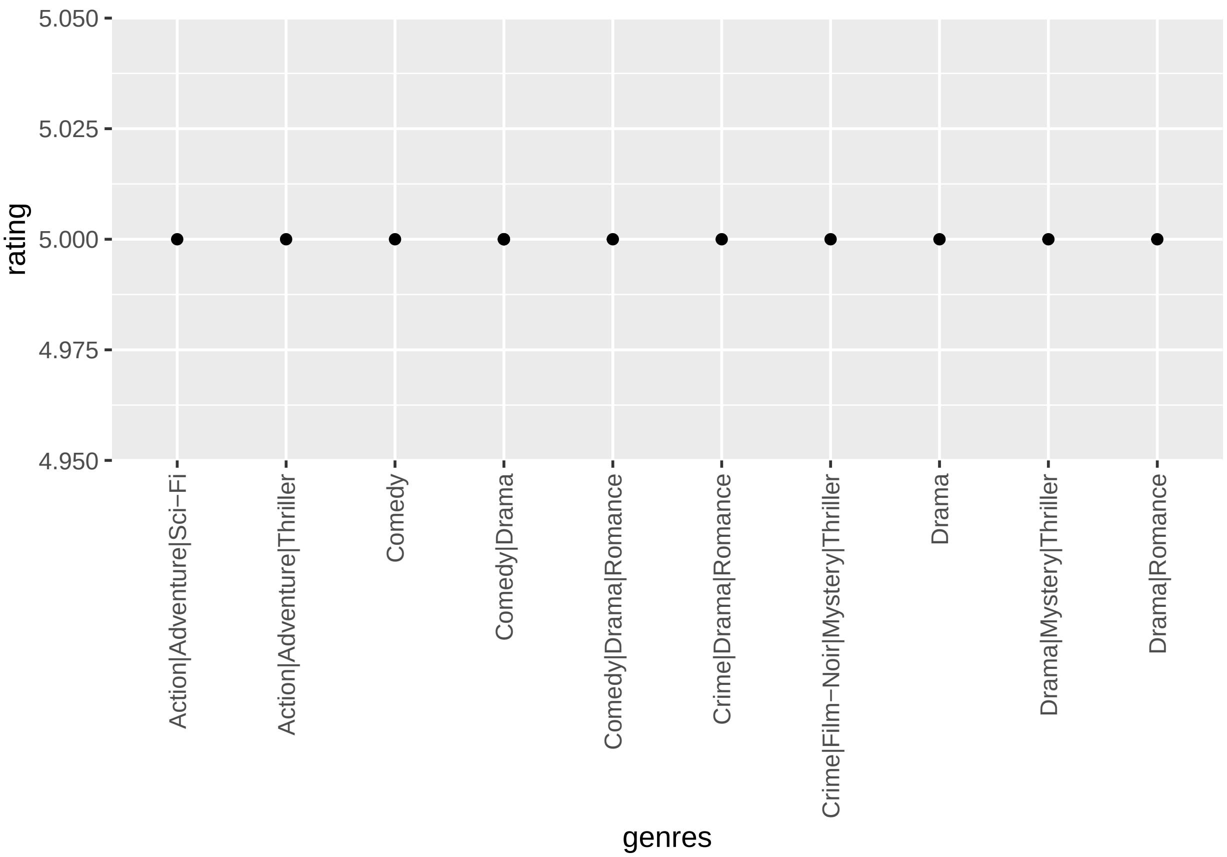
The boxplot below gives an idea of the rating for every genre. Notice how most of the genres have the same median rating. We do observe a higher tendency for Drama and Film-Noir and a lower tendency for Children and Mystery however.



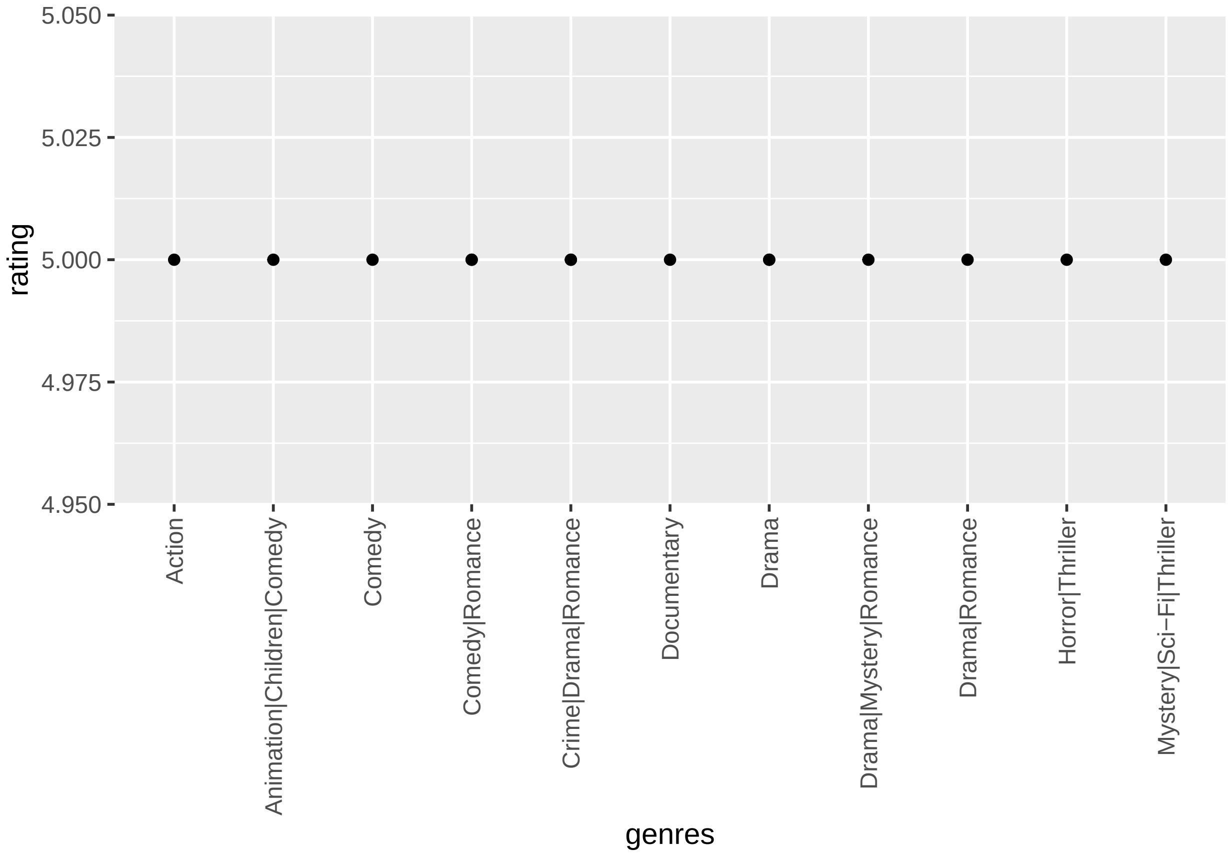
Now let's look at the users. When viewing the edx_train_ura data set, we see that userId 644 has a total rating of 6, and he's average rating is 5. That translates this specific user gives 5 stars to all movies he watched. We observe the same thing for userId 7365 and 48518. Each of them has a total rating number of 12 and 14 respectively with an average rating of 5. Even though all three of them mostly rated action movies, they also gave a 5 star rating for other movie genres.


5.050
5.025
5.000
4.975
4.950
rating
Action|Adventure|Drama
Action|Adventure|Sci−Fi
Action|Drama|War
Action|Horror|Sci−Fi|Thriller
Adventure|Drama|Film−Noir|Sci−Fi|Thriller
Comedy|Drama|Fantasy|Romance
genres

userId 644

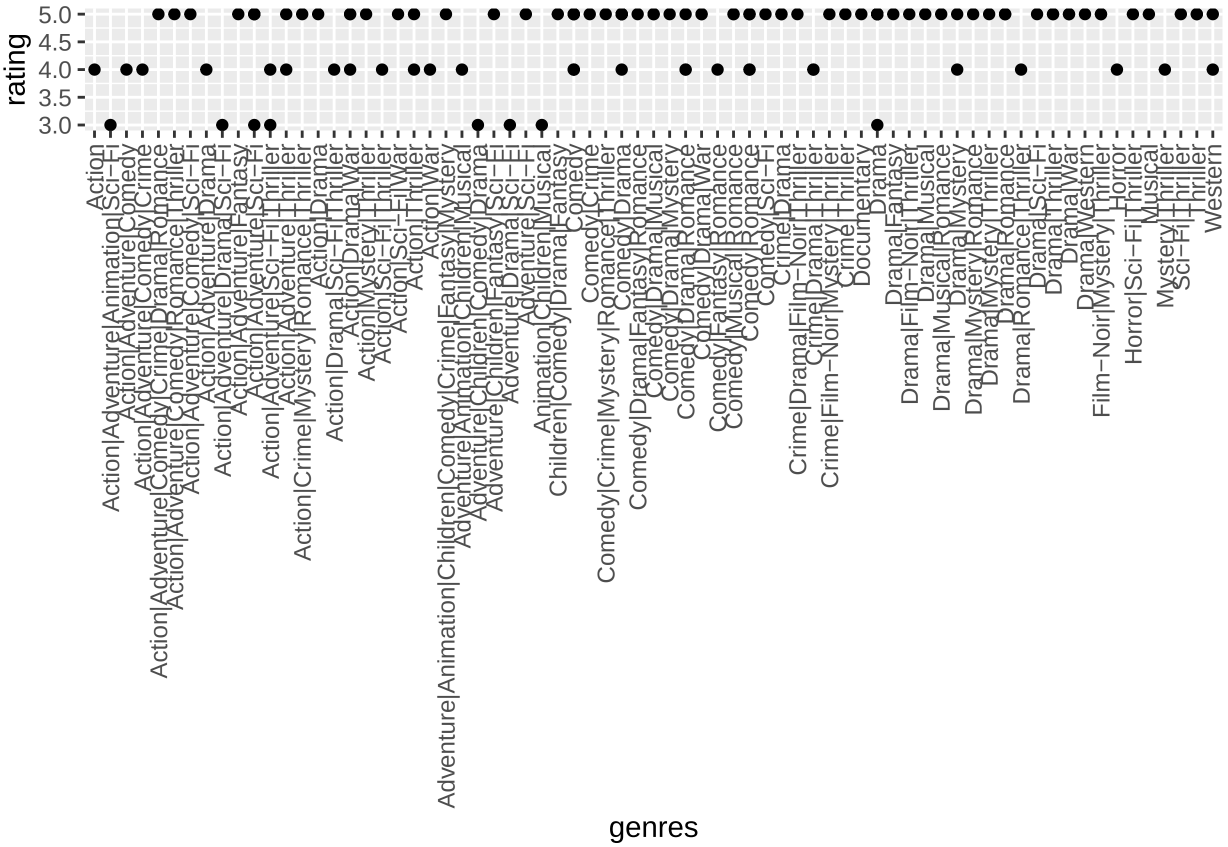

5.050
5.025
5.000
4.975
4.950
rating
Action|Adventure|Sci−Fi
Action|Adventure|Thriller
Comedy
Comedy|Drama
Comedy|Drama|Romance
Crime|Drama|Romance
Crime|Film−Noir|Mystery|Thriller
Drama
Drama|Mystery|Thriller
Drama|Romance
genres

userId 7365

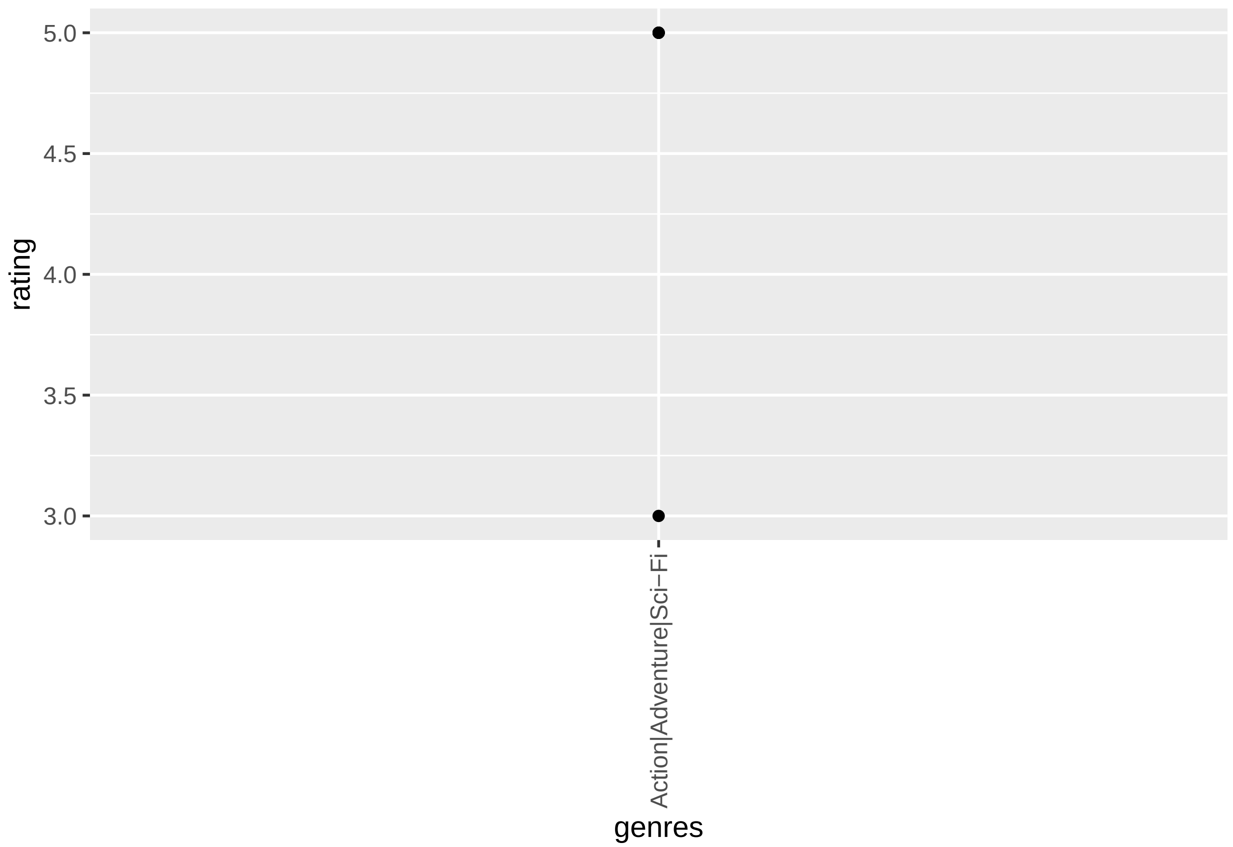


userId 48518

We are going now to search for userIds with an average rating higher than 4.6 and a total rating higher than 100. And we found only one user which userId is 32463. He has a total rating number of 131 and an average rating of 4.66. Looking at the plot, we would have thought he highly rated particular genres of movies, but this user not only has rated all the genres of movies, but had also gave them a very high rating.



Just as we said earlier, the genres of the movies by themselves cannot determine the ratings cause every movie is unique. Another thing we note here is that the user rating levels are unexpected. A user can rate several movies of the exact same genres differently. We can see an example of that for userId 32463 in the plot below. As we can see, userId 32463 rated 2 distinct movies of the same "Action|Adventure|Sci-Fi" genres at 2 different levels. We confirm that the genres doens't affect the rating that much but can be taken in account.



Basically, a movie viewer rates a movie depending on his point of view and how the movie performed overall. So it should be possible to predict the movie ratings with just the userId and movieId. We will analyse that with the proportion of variance and the training models on the data set.

We will not use the title variable as it is identified by the movieId variable. Also, as explained above, we will first calculate the PCA with just the userId and the movieId. We create the pca data set and saving it to edx_train_pca.

```
## Importance of components:
##                              PC1       PC2
## Standard deviation     2.059e+04 8928.5500
## Proportion of Variance 8.417e-01    0.1583
## Cumulative Proportion  8.417e-01    1.0000
```

When observing the summary of it, we see that we obtain 100% variability with just 2 principal components. This means, we can train our models with just the userId and movieId.

Results

We trained with the glm, knn, lda, qda, and rf models.

Here are the followings: We first use the userId and movieId predictors. With glm model, we obtain an RMSE of 1.060 with glm on the edx_test set.

```
## [1] 1.060368
```

The rf model could not perform, the accuracy metric values are missing according to the error messages.

The lda & qda training models had difficulty to perform, it generated errors.

When using most of the predictors, we accounted issues on the physical memory of the machine Error message states:“cannot allocate vector of size xx.x Gb” Computer slows down several times or training takes an eternal time with knn model process, whatever the number of predictors.

```
## [1] 1.061179
```

Finally, the glm model is the only one with which the training model can be completed and the RMSE has given a result of 1.061 on the validation set.

Conclusion

We’ve analysed the data and found that userId and movieId alone can help predict the movie ratings. Unfortunately, the entire training process could not perform properly because of computer memory fault. The majority of models could not performed as expected, with the glm being the only one acheiving the task but could not obtain an RMSE below 0.9. There is something else that can also influence the rating. It’s the age of the user. Movie viewers turn to have different point of view with age, knowing that some movies are rated for a specific age group, so the movie viewer may have different rating choices. But that will be something we can work on other movie rating projects.