MovieLens project for PH125.9x - Data Science

Wojciech Pribula

2021-05-07

1 Introduction

Main objective of this project is to build model for movies' rating prediction. This can be useful for recommendation system similar as is implemented on Netflix where Netflix wants to be able to predict if user may like specific movie.

Available training data contains 9000055 individual ratings. Each rating has, user ID, movie ID, movie title (with year included), movie genres (may have more than one), time stamp of a rating and finally rating itself. These are data which are usually available in each movies database and it should be determined if these are enough to do some predictions.

Example of data:

```
head(edx)
```

```
##    userId movieId rating timestamp                 title year
## 1:      1     122      5 838985046              Boomerang 1992
## 2:      1     185      5 838983525               Net, The 1995
## 3:      1     292      5 838983421               Outbreak 1995
## 4:      1     316      5 838983392               Stargate 1994
## 5:      1     329      5 838983392 Star Trek: Generations 1994
## 6:      1     355      5 838984474       Flintstones, The 1994
##                          genres
## 1:                Comedy|Romance
## 2:         Action|Crime|Thriller
## 3:  Action|Drama|Sci-Fi|Thriller
## 4:       Action|Adventure|Sci-Fi
## 5: Action|Adventure|Drama|Sci-Fi
## 6:       Children|Comedy|Fantasy
```

Training data are named edx and these data should be used for training. There is second collection with 999999 entries too, these data can't be used in the training and should used only for validation of predictions.

Quality of prediction should be evaluated with use of RMSE (Root-mean-square deviation) when we want to have as smallest number as possible.

$$RMSE = sqrt(\frac{\sum_{i=1}^{N}(x_i - \hat{x}_i)^2}{N})$$

N ... number of observations

x_i ... original value

$\hat{x}_i$ ... predicted value

2 Training data analysis

Data set needs to be analyzed first from point of view of each value and their relationships. Movies are easier as this topic is well known and it can be said, what should have influence on prediction, based on personal experience.

2.1 First assumptions about method

Some method should be chosen and analyzed on the beginning of this project, if it is suitable or not. If it is not then other method must be analyzed.

It seems that all predictors have some influence on the rating. It is well known that some movies are better and some are not that good so individual movie rating should be strong predictor. It is expected that some individual preferences should play role too and that different users prefer different genres.

It can be assumed that predicted rating is just most common rating (average of all ratings) plus some bias. Bias can be one or more, in this case it can be expected to have more biases, one for each predictor.

Following statements above first estimation of model may look like this:

$$Y = average + bias_{movie} + bias_{user} + bias_{year} + bias_{usergenre} + bias_{genre}$$

$average$ ... average rating for all data

$bias_{movie}$ ... bias for individual movie - better OR worse movies

$bias_{user}$ ... user bias - represents tendency of user to be picky or generous

$bias_{year}$ ... bias for year of movie release

$bias_{usergenre}$ ... this should represent users genre preference

$bias_{genre}$ ... some genres may be rated better as they may be more popular

2.2 Average

Distribution of ratings and average should provide some basic idea about data set.

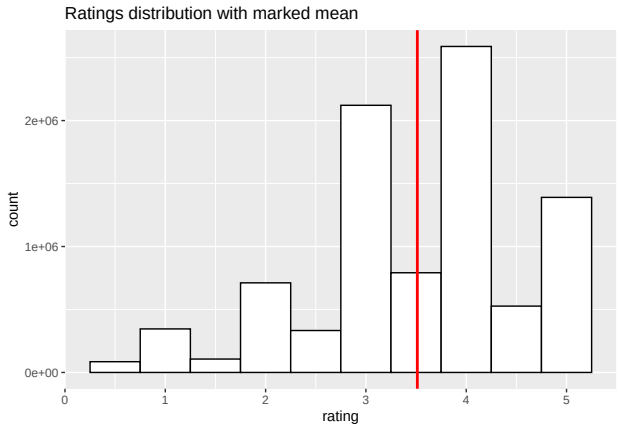

Figure 1: Ratings distribution with marked mean.

If rating should be predicted as global average then rating is 3.5124652 and RMSE for this is 1.0612018. This is value which should be beaten by prediction model as this is the simplest prediction.

2.3 Movie bias

Distribution of movie ratings suggests that most movies are rated around the average, but there is big group of movies rated better and worse, so this should be very good predictor.

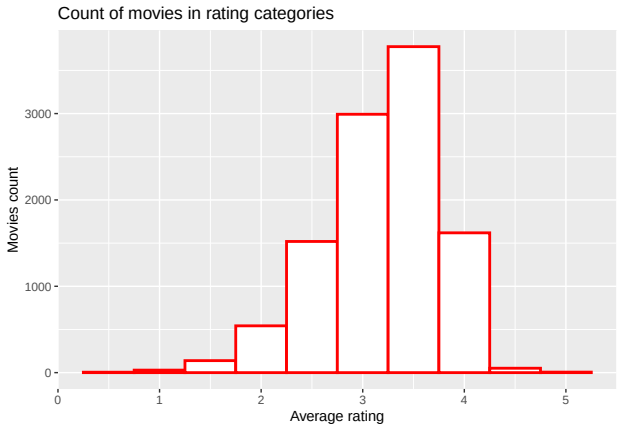

Figure 2: Count of movies in rating categorie.

Another factor which should be taken in count is fact that very high rated movies have only few ratings what may suggest that these are rather niche movies picked by users who already like this kind of movies, so these movies ratings may not be accurate, when generalized on bigger population. It can be similar for bad movies, however these are naturally less watched so less rated, so this is not visible in the plot.

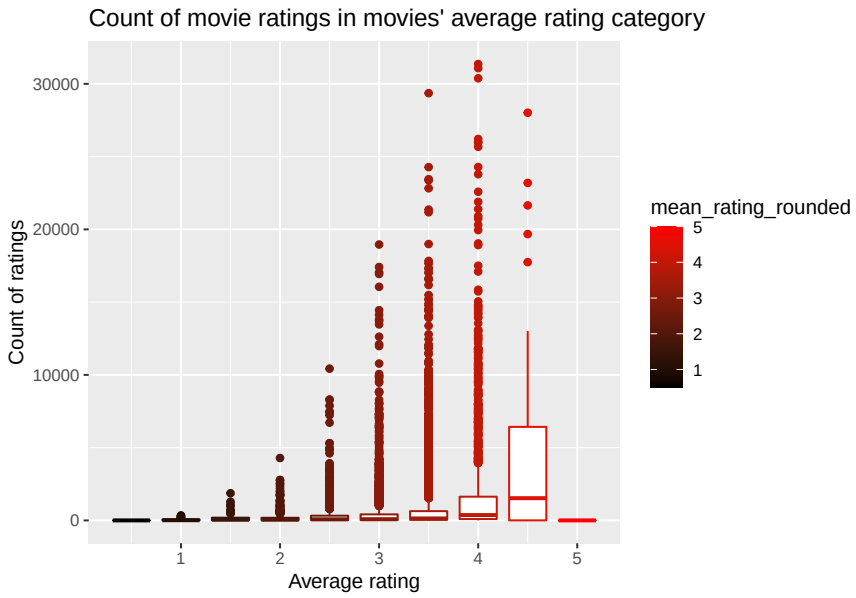

Figure 3: Count of movie ratings in movies' average rating category.

2.4 User bias

User bias can be seen in users average rating distribution. It can be seen that most users rate around average but some users may have strong bias what makes it good predictor.

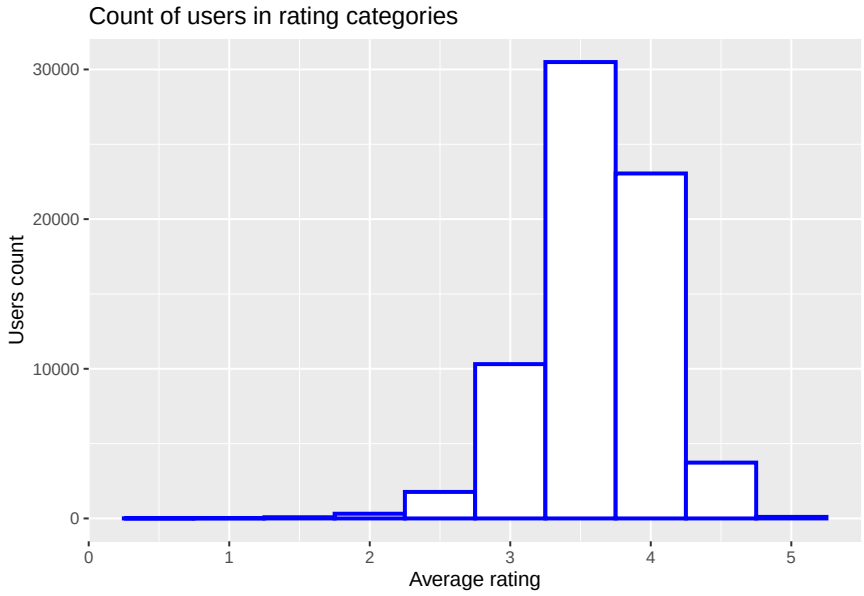

Figure 4: Count of users in rating categories.

What can be expected is that users with only few ratings are not well represented by their average rating as there is not enough data to tell anything about their preferences or bias. It is confirmed by following plot where users with less ratings have wider spread of ratings, with more ratings per user ratings incline to average.

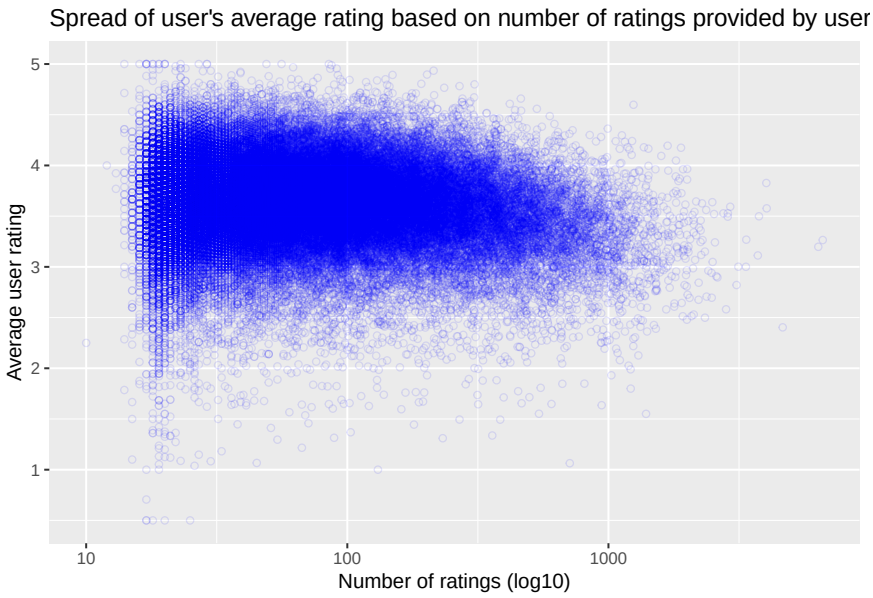

Figure 5: Spread of user's average rating based on number of ratings provided by user.

2.5 Year bias

Year when movie was released can have some influence on the rating, this can be confirmed using following plot.

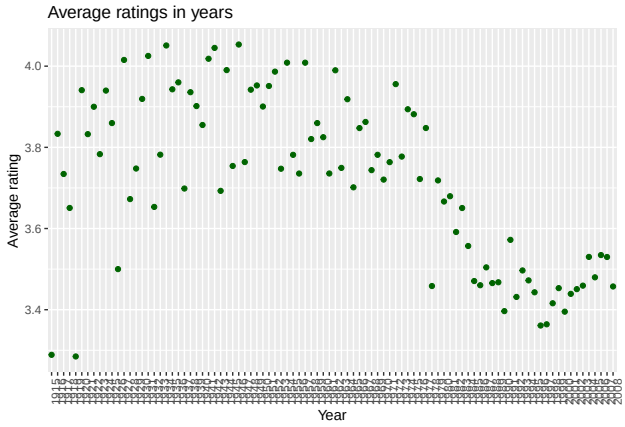

Figure 6: Average ratings in years.

Clear dependency can be be seen, people probably prefer newer movies, or old movies lovers are more picky and rate more harder. This would be great addition to prediction model.

2.6 Genre Bias

First what must be done is separation of genres from list of genres for individual movies. Then genres ratings can be inspected in the box plot.

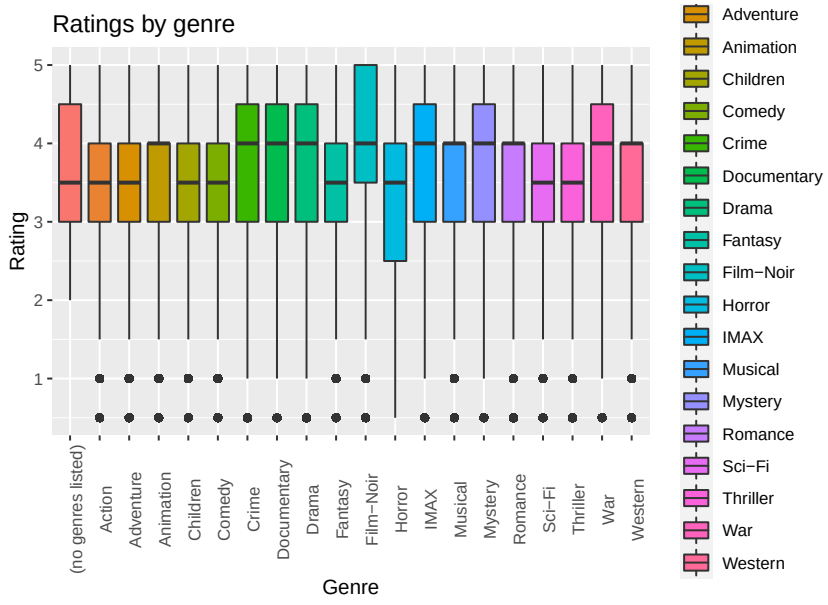

Figure 7: Ratings by genre.

It is clear that some genres receive different ratings on average than others and taking in count genre can enhance final prediction. Prediction should be careful as data do not provide the same size of sample for all genres.

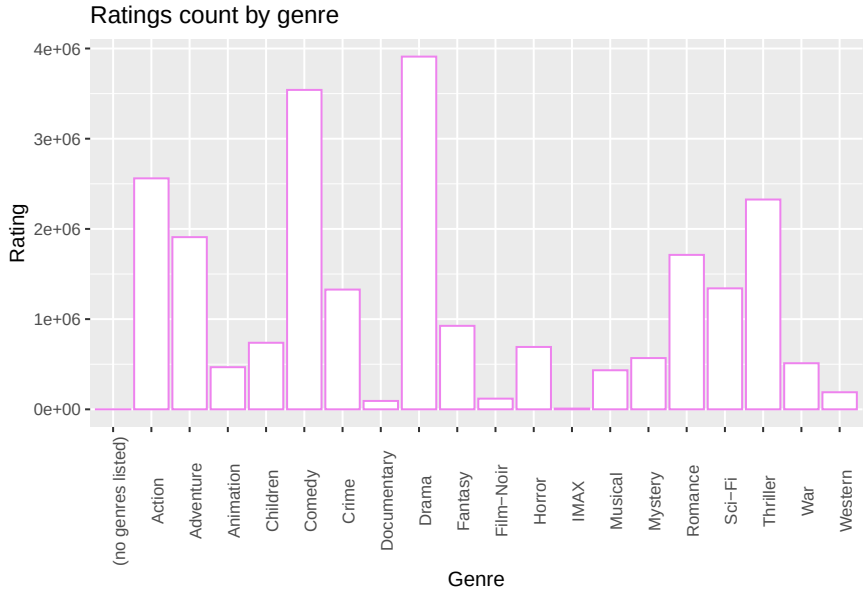

Figure 8: Ratings count by genre.

2.7 User and genre bias

This should be predictor based on combination of user and genre as experience suggests that some users may like some genres more than other. This can be seen on the plot of selected five users with biggest number of ratings.

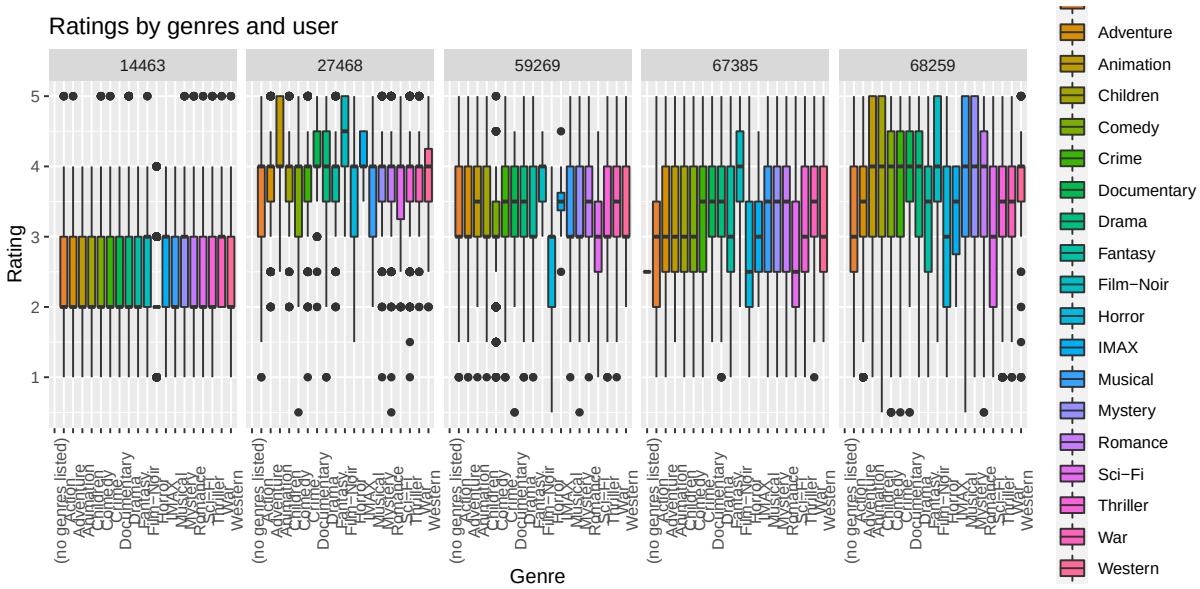

Figure 9: Ratings by genres and user.

Genre preferences can be clearly observed for all users expect the first one which is very consistent in ratings across genres. Second and fifth user present very clear overall individual bias too what is supporting claim from user bias chapter.

3 Final data model and training

There is clear bias for all described categories and all of them can improve prediction. What must be taken in count is fact that some movies have more genres than one, so model should be corrected.

$$Y = average + bias_{movie} + bias_{user} + bias_{year} + frac1N_g \sum_{i=0}^{N_g} bias_{usergenre,i} + frac1N_g \sum_{i=0}^{N_g} bias_{genre,i}$$

$average$ ... average rating for all data

$bias_{movie}$ ... individual movie influence - better OR worse movie

$bias_{user}$ ... user bias represents tendency of use to be picky or generous

$bias_{year}$ ... predicted value

$bias_{usergenre}$ ... this should represent users genre preference

$bias_{genre}$ ... some genres may be rated better as they may be more popular

N_g ... number of genres

3.1 Biases training

Bias should be calculated as average of ratings differences from global average for individual categories. As there is influence of small groups of ratings average can be hardened by α parameter which can be tuned.

$$Bias_{cat} = \frac{1}{\alpha + N} \sum_{i=0}^{N} (rating_i - \mu)$$

α ... tuning parameter

N ... number of ratings for category

$rating_i$ ... individual ratings

μ ... global ratings average

Parameter $\alpha|$ should be tuned with use of RMSE function and results of tuning should be presented in the result schapter. Calculation of RMSE is possible as train data can be divided into training and test sets. Data set edx can be divided 1:4 (test:train).

3.2 Cross checking (Ensemble model)

As division of edx is random cross checking can be done with applying training process multiple times. Final model may be calculated as combination of multiple models:

One possibility, which is used in this project, is average from all individual models.

$$M = \frac{1}{N} \sum_{i=1}^{N} M_i$$

N ... number of models

M_i ... individual models

4 Results

Final precision for current data is:

$$RMSE = 0.8513403$$

However if three training runs are compared some issues can be observed.

Three training runs were run. Each include 5 individual models. Different random seed was used to produce different train and test sets from edx. Here are RMSE for all models by training run. Final RMSE of ensembled model is marked with a line. It can be seen that ensembled model usually enhances individual models.

Figure 10: RMSE for 5 models in mulliple trainings - lines show final RMSE of ensembled model.

This is how it looks for 1 to 5 ensembled models. It can be see that 5 models is usually good prediction and that final model is better when 5 models is ensemled, however for model 3 it is clear that adding more and more models make predictions worse, it is caused by adding worse models, if order of models adding would change from 5 to 1 the line would be declining.

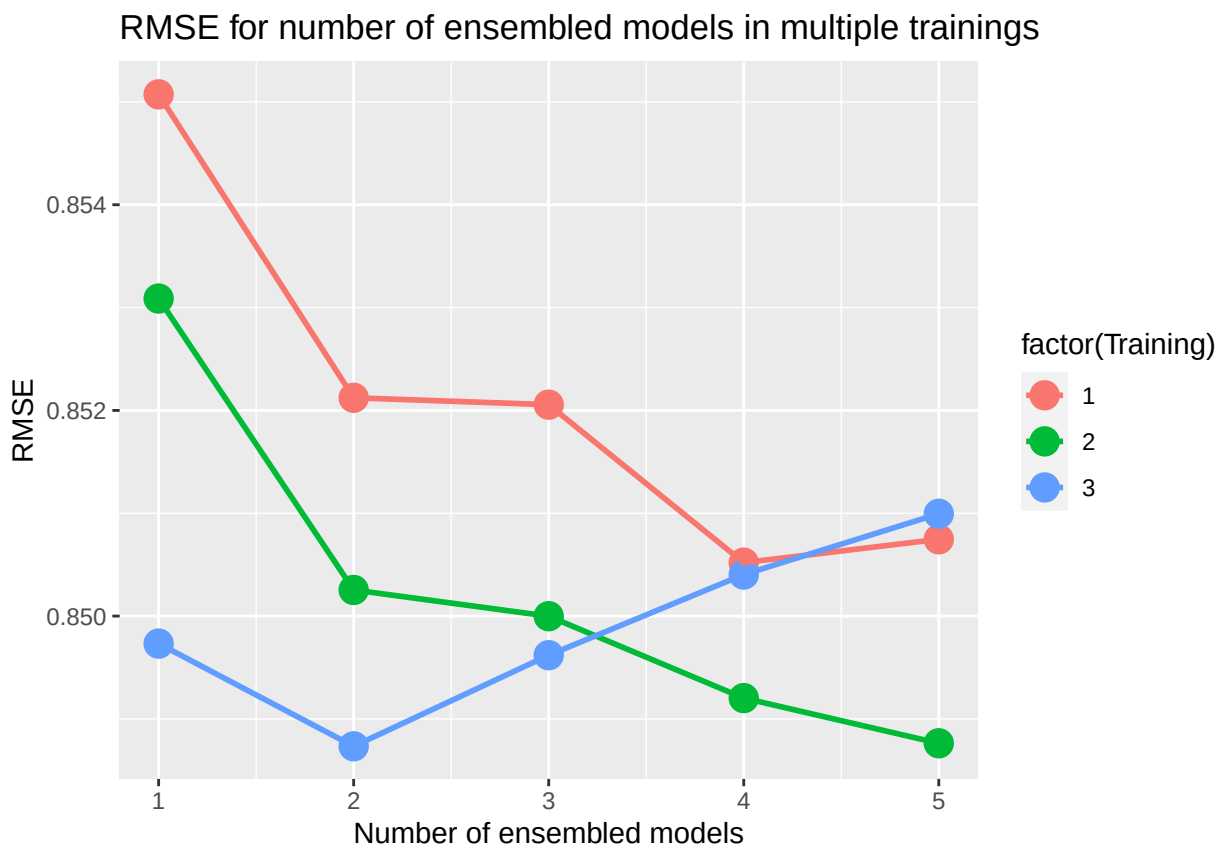

Figure 11: RMSE for number of ensembled models in multiple trainings.

It is clear that 5 models may perform worse than just one, however it is not clear which model of these five is the best, so making average of multiple models is safer than keeping just one.

4.1 Alpha tuning results

Alpha tuning results can be examined in following plots. Alpha boundaries were configured based on experiments.

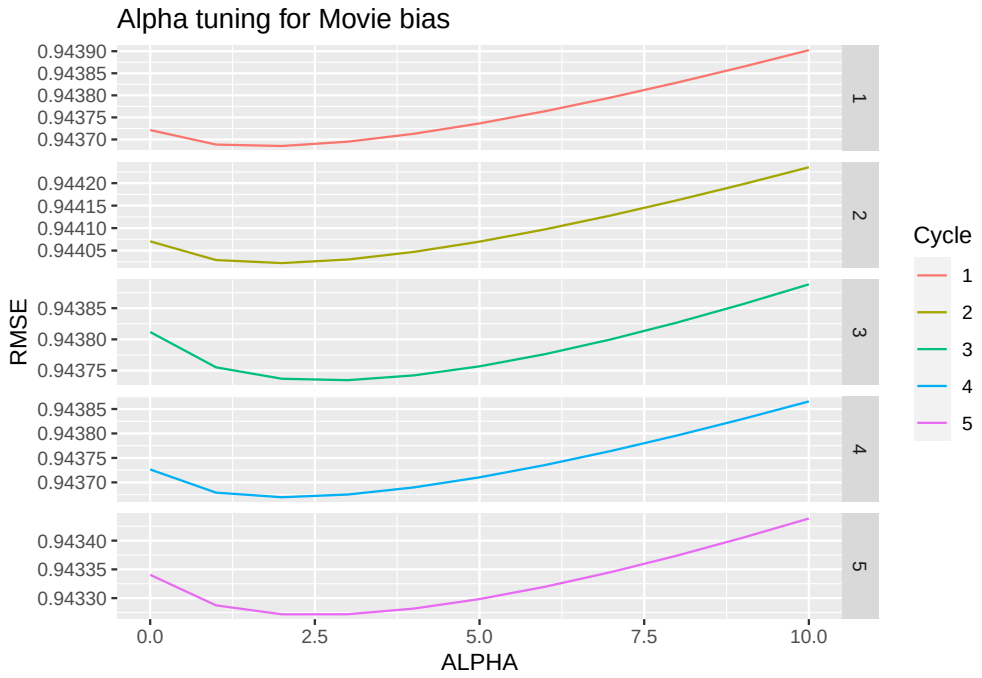

Figure 12: Alpha tuning for Movie bias.

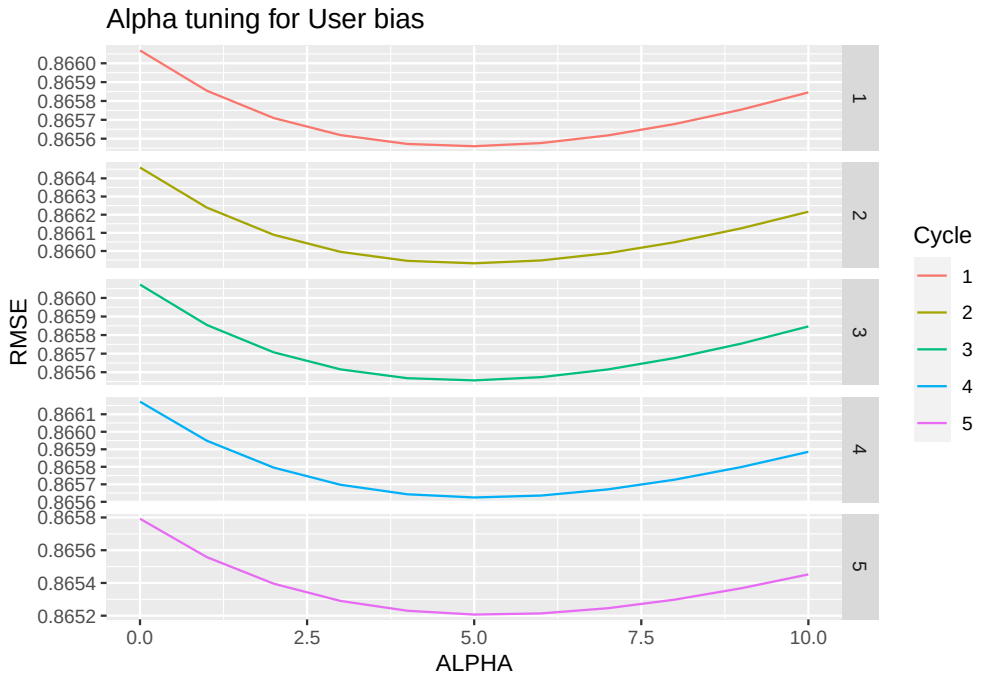

Figure 13: Alpha tuning for User bias.

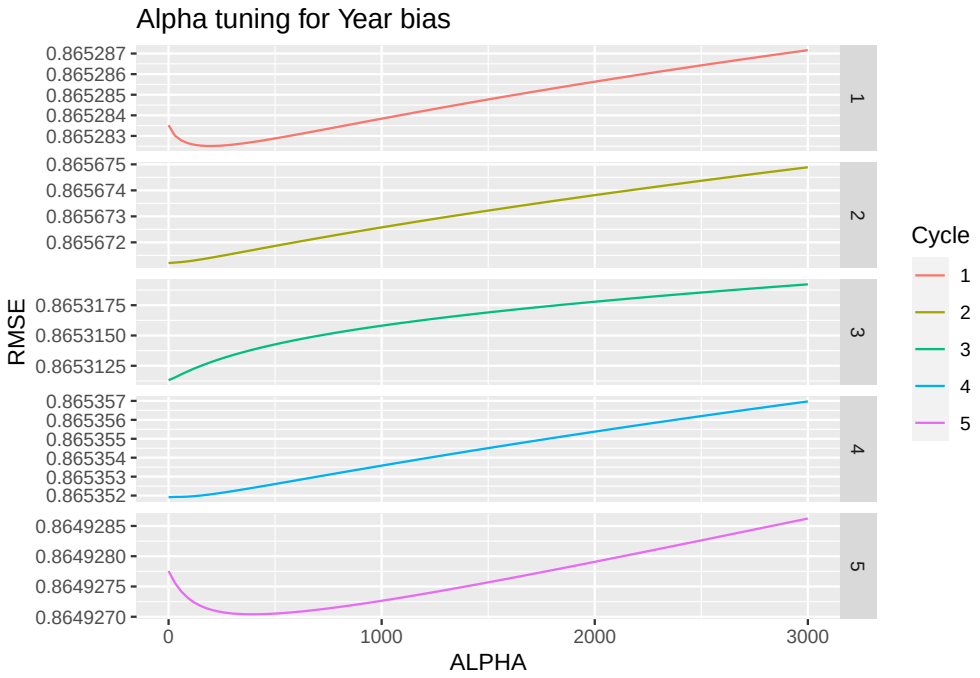

Figure 14: Alpha tuning for Year bias.

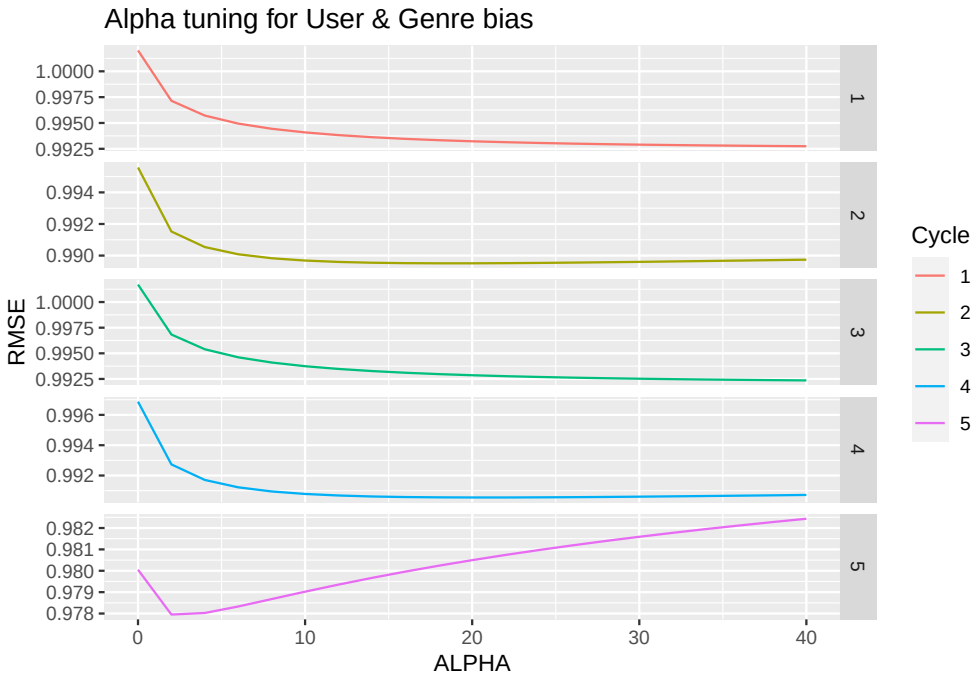

Figure 15: Alpha tuning for User and Genre bias.

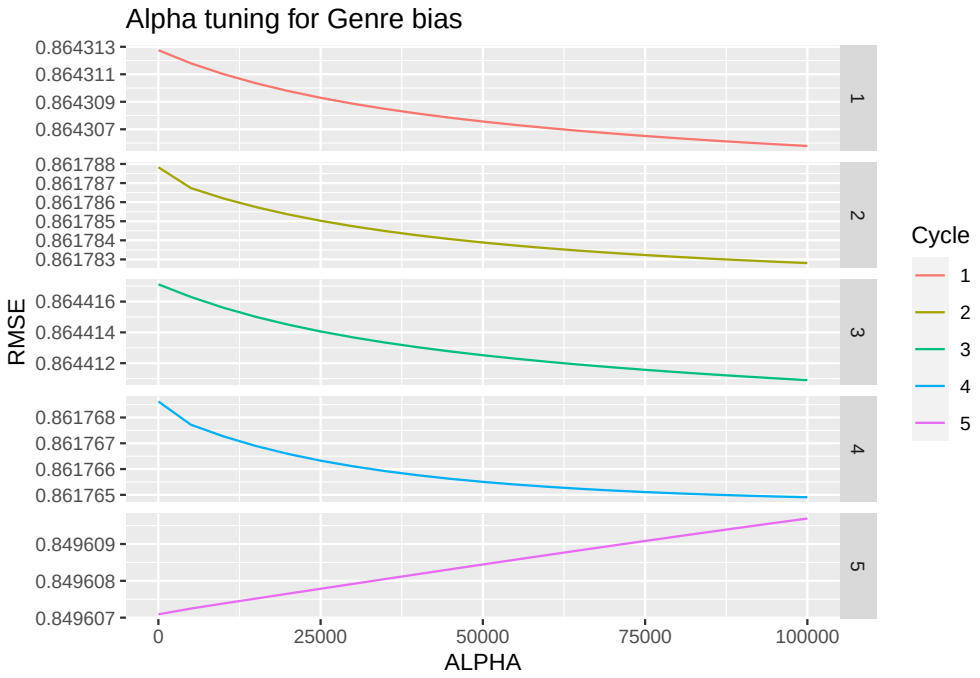

Figure 16: Alpha tuning for Genre bias.

5 Conclusion

Final precision is $RMSE = 0.8513403$ however result could be better if more individual models would be trained and added to the ensembled model. It is always some kind of compromise between time for training and results.

It is possible to come with some algorithm which couldbe working with bigger number of individual models, would be sorting them by precision on training data and use only top performing algorithms for final model. However this would require much more training time, however this is something what could be added as enhancement to the current project.

Tuning of α parameter follows some predefined boundaries which were configured based on experiments, it would be interesting to do more investigation on this and observer influence of alpha tuning on final model performance.