Summary

A subset of MovieLens dataset (prepared by the HarvardX team) has been analysed and used to build a predictive model which predicts movie ratings. The data was imported, evaluated for relevant predictors and fitted.

The final model includes individual rating, movie and age of movie as predictors. This model resulted in a residual mean squared error of 0.825 – which is deemed adequate according to the marking rubric of the relevant assessment at hand.

Introduction

The MovieLens data was collated by the GroupLens Research group and widely used as a practice data set for education purposes in statistical modelling and data science.

Note that for this assignment, the HarvardX team has subsetted the data into a smaller data set of:

- Edx dataset of 900,005 observations.
- Validation dataset of 999,999 observations.

Both datasets contain the variables:

- User ID: Individual user ID for raters. Note that this is important as most of the raters gave more than one rating.
- Movie ID: Corresponding ID for movies.
- Rating: Rating given by specific User ID and Movie ID – this is the outcome that the model is predicting (i.e., dependent variable).
- Timestamp: Relevant timestamp data which outlines the time the rating was given. Note that this is further processed and the age of each movie is calculated from this variable.
- Genres: Relevant associated genre for each movie.
- Title: Relevant title for corresponding movie.

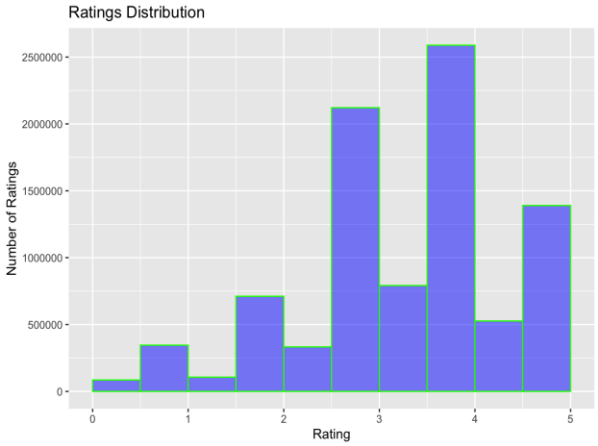
NA observations were checked for. Data were then examined to determine the relevant factors deemed significant for modelling.

Note that throughout the entire analysis of data and initial model building, only the Edx dataset was used. The validation dataset was only used after the model was built for RMSE evaluation purposes.

Analysis: Analysing Data to Determine Relevant Predictors

1) Understanding the properties of 'ratings'

To further understand the data, the distributive properties of ratings were examined.



The distribution of ratings does not look normal on prima facie; however, this is due to the .5 ratings being given less frequently than round ratings.

Note that the distribution of both round and .5 ratings follow the same pattern:

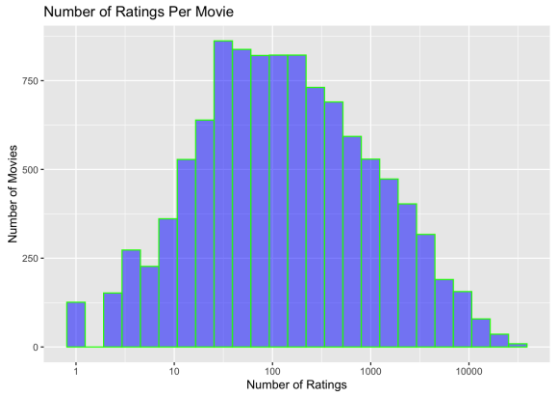
- There is a rising trend in numbers of round ratings from 0 and this peaks off at '4' and decreases at '5'.
- There is also a rise of number of .5 ratings from 0, peaking off at 3.5 and decreasing after that.

Note that both these indicates that there is nothing significantly abnormal about the ratings distribution – and hence, no further transformation of data will be conducted.

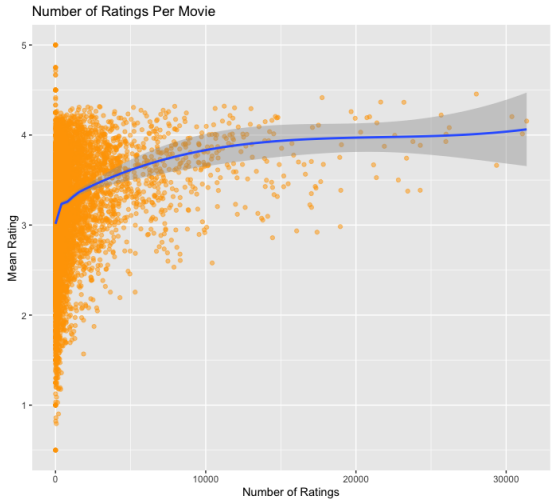
Through this, it can be inferred that each movie has a different mean rating – and hence, movie ID is deemed a significant variable to be included in the model.

2) Examining whether the Number of Ratings per Movie Effects Ratings

As expected, not all movies are equal – some are block busters and others are indie films.



A further plot examining the relationship of the number of ratings and average ratings for each movie was conducted.



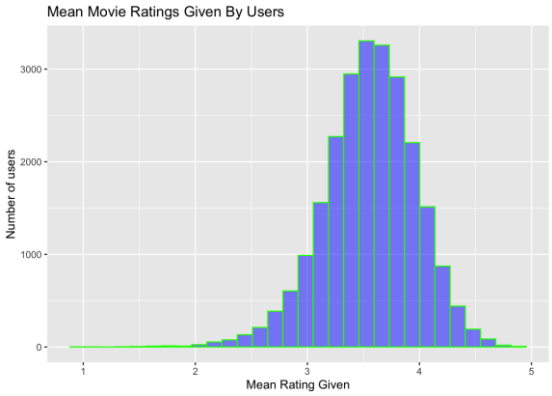
The loess line plotted here indicates that there is a larger variation in mean ratings in movies that are not rated a high number of times.

Regardless, the plot here indicates that the relationship between number of ratings for each movie, and, average ratings is weak.

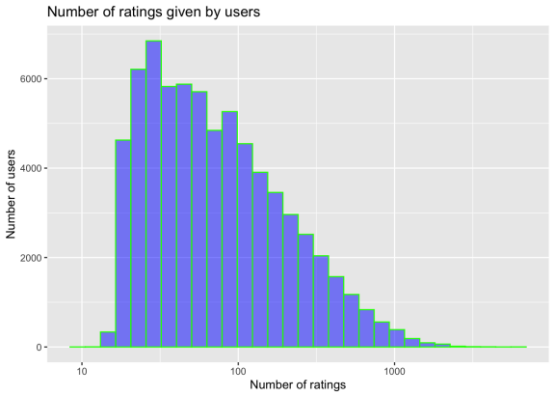
Hence, the decision to not include this model was made.

3) Examining User Effects in Ratings

The number of ratings given by users distribution, and, mean movie ratings distributions have been plotted to examine for user effects.



The mean movie ratings distribution graph indicates that users give different mean ratings – this is to be expected as most users have different baseline for ratings (i.e., rating bias).



A further graph plotted indicated that the majority of raters give below 100 movie ratings.[1]

This is to be expected as not all raters is as interested or passionate in giving movie ratings.

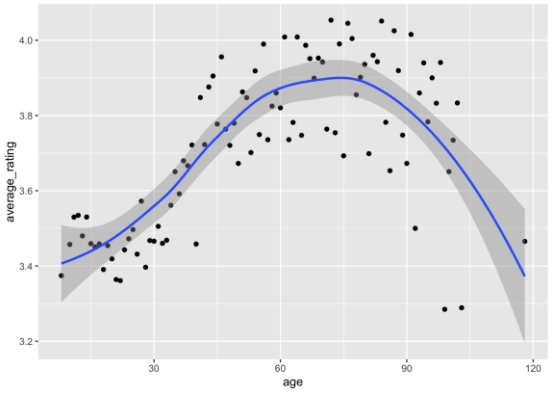
Both these indicate that user effects are deemed significant and hence, will be included in the model.

[1] Note that the x axis is transformed with log 10 for presentation reasons.

4) Examining Age Effects of Movies on Ratings

The age of movie was calculated. This was done by parsing the time stamp variable. The age of each movie was then calculated.

A further plot examining the relationship between the age of movie and average rating was plotted.



Note that there is a relationship between age and average ratings – hence, the age was deemed a significant predictor and decided to be included in the model.[2]

Preliminary Model and Fit Results

Based on the prior analysis, the predictors decided to be significantly relevant to be included in the model are age, User ID and movie ID.

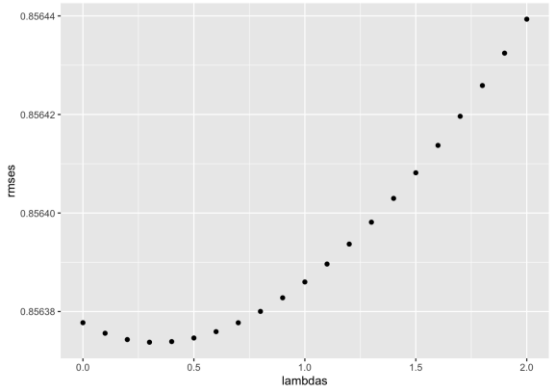
A linear model was fitted and indicated an adjusted R squared of 0.014. Whilst the R squared is low, it is important to note that RMSE is the evaluating benchmark here.

The purpose of this fit is to determine whether the predictors chosen are statistically significant relevant – the fit indicated that all predictors are significant at $p<.05$ and hence will be included in fitting the final model.

[2] 2018 was used as a baseline as the data used was released in 2018 (as per Group Lens' website). Hence, a movie released in 2000 would be coded as '18' years old.

Modelling: Fitting and Regularisation

The model is then fitted with regularisation. The parameters were decided based on the lowest RMSE achieved on the edX dataset.



Note that the optimum lambda that minimises the RMSE is 0.3

Results : Evaluation of Model

The model is then tested based on the validation dataset. The age was calculated for each movie in the validation dataset before fitting the model.

The final RMSE obtained from the model is 0.8250394.

Conclusion

The final model fitted here is:

Ratings = User + Movie + Age of Movie

All 3 predictors here have been regularised with lambda = 0.3.

Limitations

Note that the modelling here is not extensive, as the sole purpose of this assignment is to achieve an adequate RMSE.

Hence, further models could be fitted and tested for a lower RMSE – however, as the goal has been achieved in this initial model fitted, no further work was deemed done.