MovieLens

Penny Cookson

10/12/2021

MoveLens Project - Report

Introduction

Aim

This report documents an analysis of movie rating data, describes the methodology used to create a ratings prediction model and reports on the results obtained.

Executive Summary

The final model achieved a RMSE of 0.859562. This model included factors for the effect of the movie, user, and time, and also added a factor for the preference a user has demonstrated for movies of a particular genre.

Initial predictions were based on the average ratings for the movie and user.

Including time in the model provided only a small improvement in prediction.

A key factor in predicting a user's rating of a movie was to consider the ratings given to movies which were classified as having one or more of the same genres as movies previously rated by the user. The analysis looked at how the user rated movies with genres assigned and, using the difference from their average rating, determined whether the user had a preference for movies of that genre. This allows the use of their genre preferences in predicting how they will rate new movies.

It would be expected that improved predictions could be obtained if additional demographic data for the users was available, or if multiple more sophisticated machine learning algorithms could be executed on more powerful hardware, and combined to form an overall prediction. However the final model used in this analysis utilises all the provided data and provides a simple prediction with an aceptable level of error.

Source Data

The data set is sourced from GroupLens (1) which provides non commercial personalised movie recommendations. The 10M version of the MovieLens dataset was used. This data set contains 9,000,055 ratings applied to 10,677 different movies by 69,878 users of the online movie recommender service MovieLens. This is a stable benchmark data set.

It is important to note that, unlike most recommendations data sets, this data does not include any demographic information for users. The analysis therefore relies heavily on the movies each user has already rated.

The aim was to create a model to allow the prediction of ratings of movies by users, given their past movies ratings.

Key Steps

The following steps were performed:

1. Data loading
   Data was loaded from the given source and saved for future use.

2. Data quality checking
   Quality checking included empty and invalid values.

3. Data exploration and visualisation
   Visualisation of data was used to evaluate the attributes which could be used to predict ratings.

4. Model creation and evaluation.
   The initial model used the approach provided in the edX Data Science course materials. The main difference being that tuning parameters were determined from replicated sampling from the train set, rather than the test set (which was used in the course example, and which could lead to overtraining).

   This model was then extended as a result of the factors determined during the visualisation. Initially time was included, followed by the manipulation of the data to allow preferences for individual genres to be determined for each user.

   There is limited data that may assist in identifying the type of movies which a user may rate more highly, for example there is no demographic data for the users.
   It seemed likely that users would have a preference for movies of a particular genre. For example a user that generally rates movies with a FilmNoir genre highly may be likely to prefer other movies of the same genre. Users who generallly give Romances a low rating may have a pattern of this behaviour. The visualisations support the fact that users have preferences for certain genres.

   As a result, the main technique to improve the predictions was to assign each user a rating for each genre, based on their previous ratings of the genre and include this as a factor in future predictions.

   Each of the models was optimised using replicates sampling from the train set, and then evaluated against the test set.

5. Final model evaluation
   Final model evaluation was performed against the validation set provided in the course materials.
   The measurement of Root Mean Squared Error (RMSE) was used to evaluate the accuracy of the prediction model. The aim was to achieve a RMSE less than 0.86490.

Analysis

Data Source

The data provided was divided into a train data set (80% of the orginal data set), and a test data set (20% of the original data set). The test data set was used only to provide an initial evaluation of the models created.

The validation data set was used only in the final stage to assess the selected model.

Data Quality Checking/Cleaning

Structure

The data has the following structure:

Column Name	Data Type
userId	integer
movieId	numeric

Column Name	Data Type
rating	numeric
timestamp	integer
title	character
genres	character

Counts

Number of rows: 9,000,055
Number of Movies: 10,677
Number of Users: 69,878

Minimum number of Movie Ratings per User: 10
Maximum number of Movie Ratings per User: 6,616
Mean number of Movie Ratings per User: 128.8

Minimum number of Movie Ratings per Movie: 1
Maximum number of Movie Ratings per Movie: 31,362
Mean number of Movie Ratings per Movie: 842.94

It is noted that there is considerable variation in both the number of ratings per user, and the number of ratings per movie.

Data Quality

Data quality checking indicated the following:
* None of userId, movieId, title, genres, or timestamp contained null or NA values.
* userId values were all valid integers.
* movieId values were all numeric.
* timestamp values were all valid integers.
* There were no inconsistent values for title or genre for any movie.
* Ratings were all one of the following values (0.5,1,1.5,2,2.5,3,3.5,4,4.5,5).

Exploration and Visualisation

Splitting the Genres

The genres were provided as a | separated character string. This was separated into individual genre columns to allow the genre effect to be investigated.

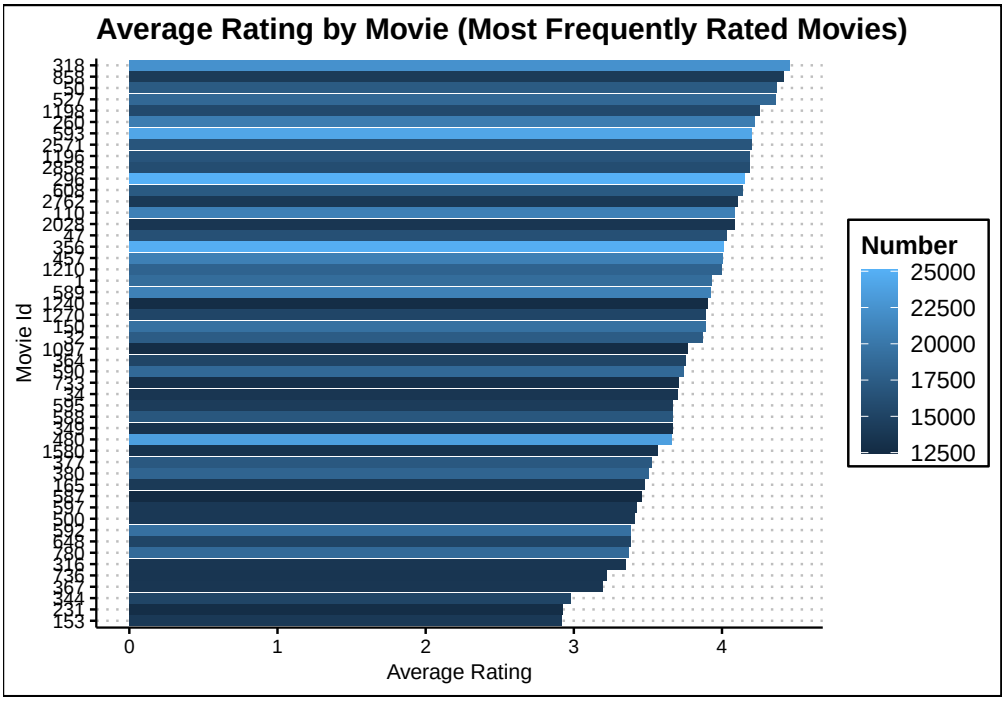
There are 21 distinct genres. The maximum number of genres for any movie is 8.

Visualisation of Factors Affecting Ratings

Note that for the following visualisations only the train data set was used.
In order to determine which attributes affect the ratings, and may be useful for prediction, the following were plotted:

Movie Effect

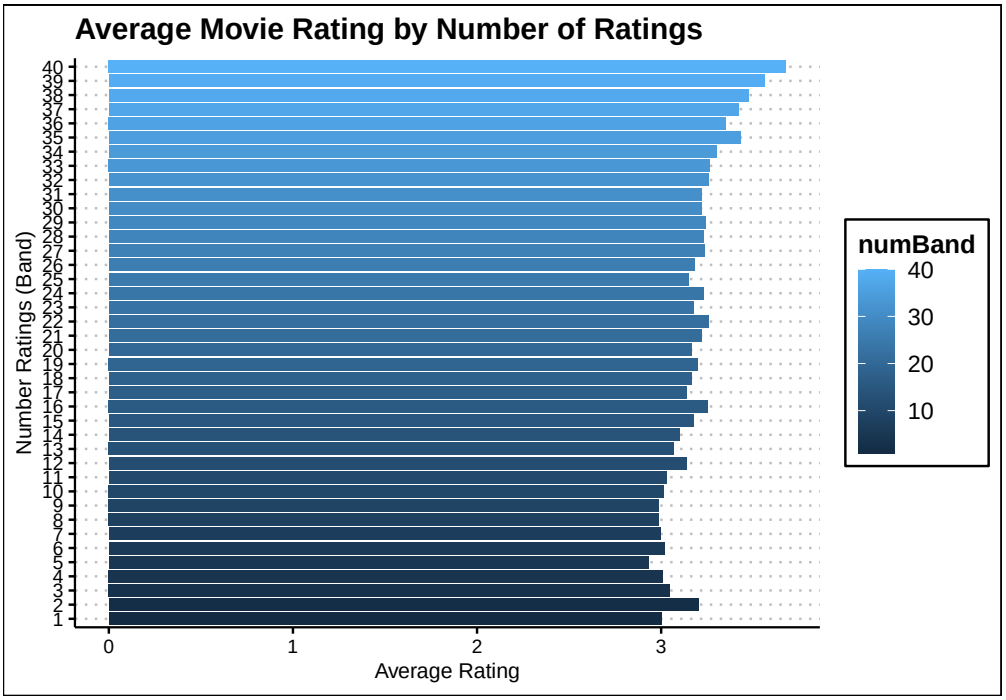
In order to investigate the effect of individual movies the focus is on on the 50 most rated movies. The average rating for the 50 most frequently rated movies was plotted.



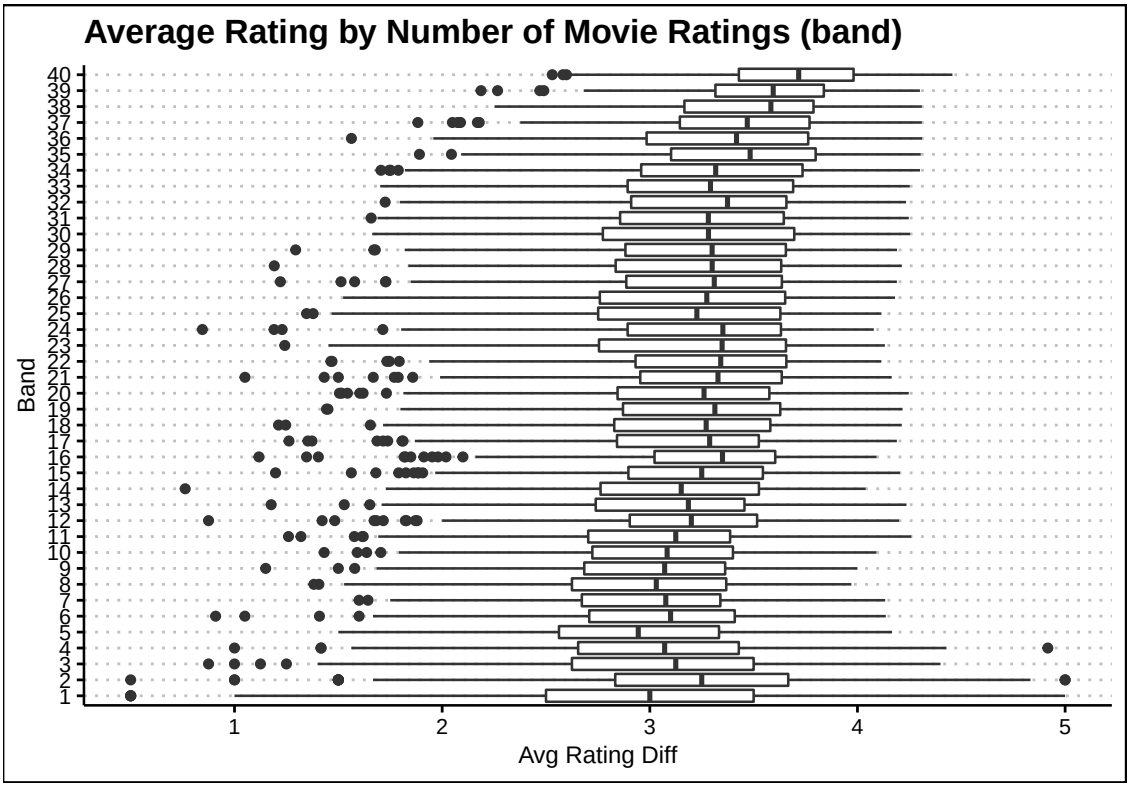
Is the individual movie a factor in ratings?

The plot indicates that there is a difference in average rating of more than 1 between the lowest and highest rated movies in this group. We can see that the movie has an influence on ratings. The colour indicating the number of rating appears scattered randomly throughout movies of various ratings so we do not expect the number of times a movie is rated to have a significant effect in this set of frequently rated movies.

The following plot investigates any relationship between the rating and the number of times a movie is rated. Since the number of times a movie is rated varies widely, this has been divided into 40 bands using a balanced number in each band.



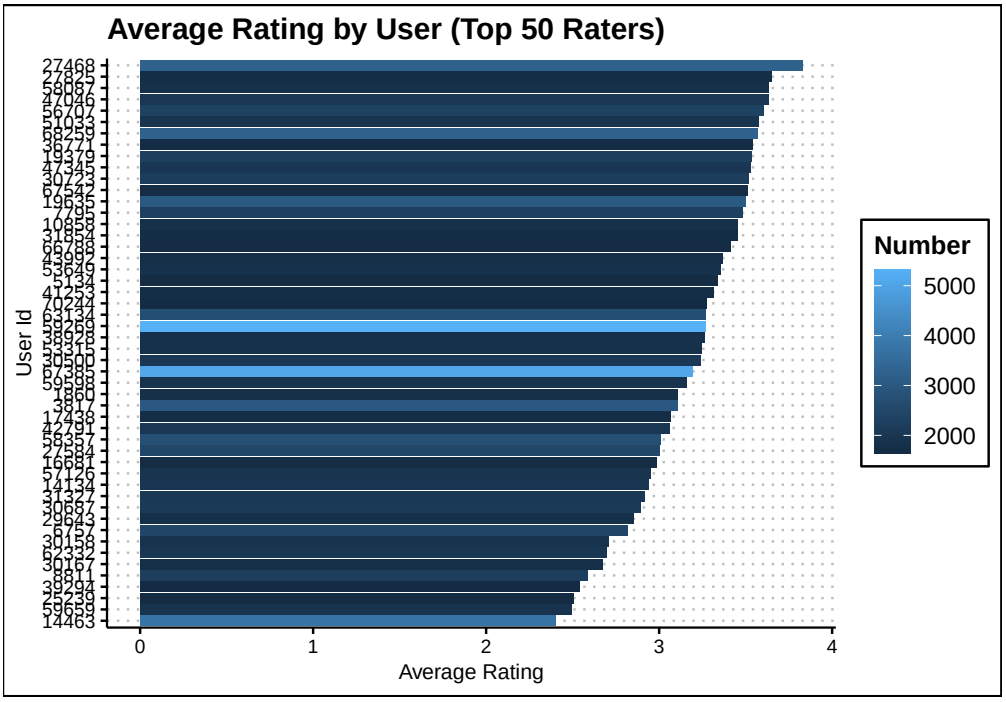
Is the number of times a movie is rated a factor in ratings?
The plot above indicates that movies that are rated more frequently have a very slightly higher rating generally. This data however looks at the mean rating. in the following plot the range of ratings these bands take is investigated, and whether there are signficant outliers.



What is the range of ratings for movies in a band, and are there signifiant outliers?
The plot clearly indicates that the movies that are not rated so often have a wider range of ratings, and significant outliers. As a result of this finding the model will need to regularize the predictions to reduce the variability of these low rated movies.

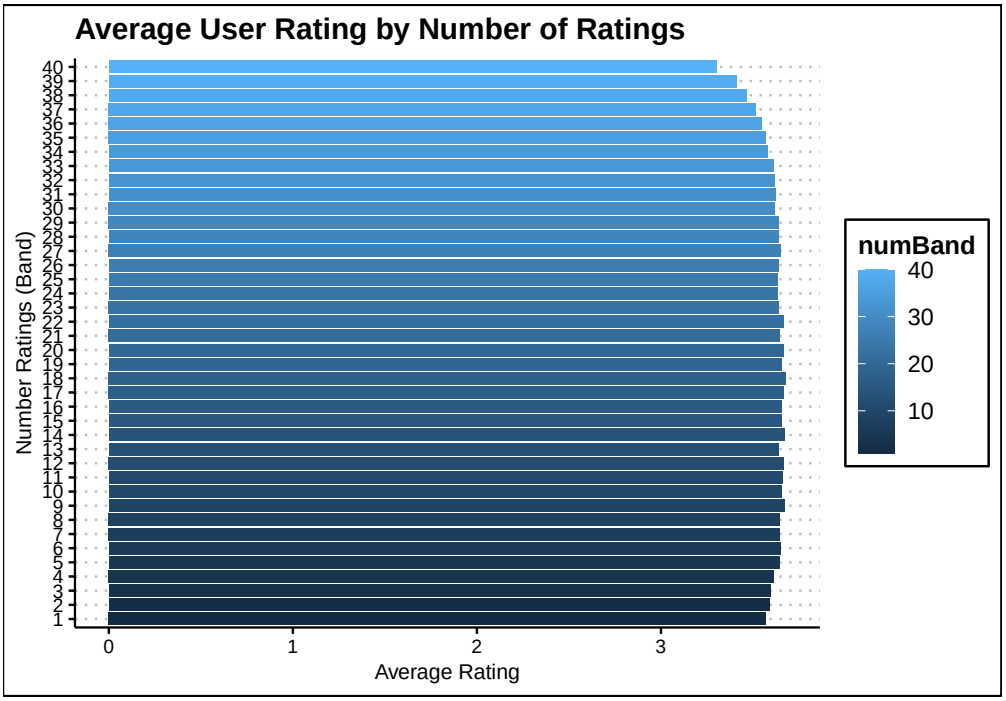
User Effect

The next section performs a similar investigation for individual users. In order to investigate the effect of individual users the focus is on the 50 users with the most ratings. The average rating for most 50 most frequently rating users was plotted.

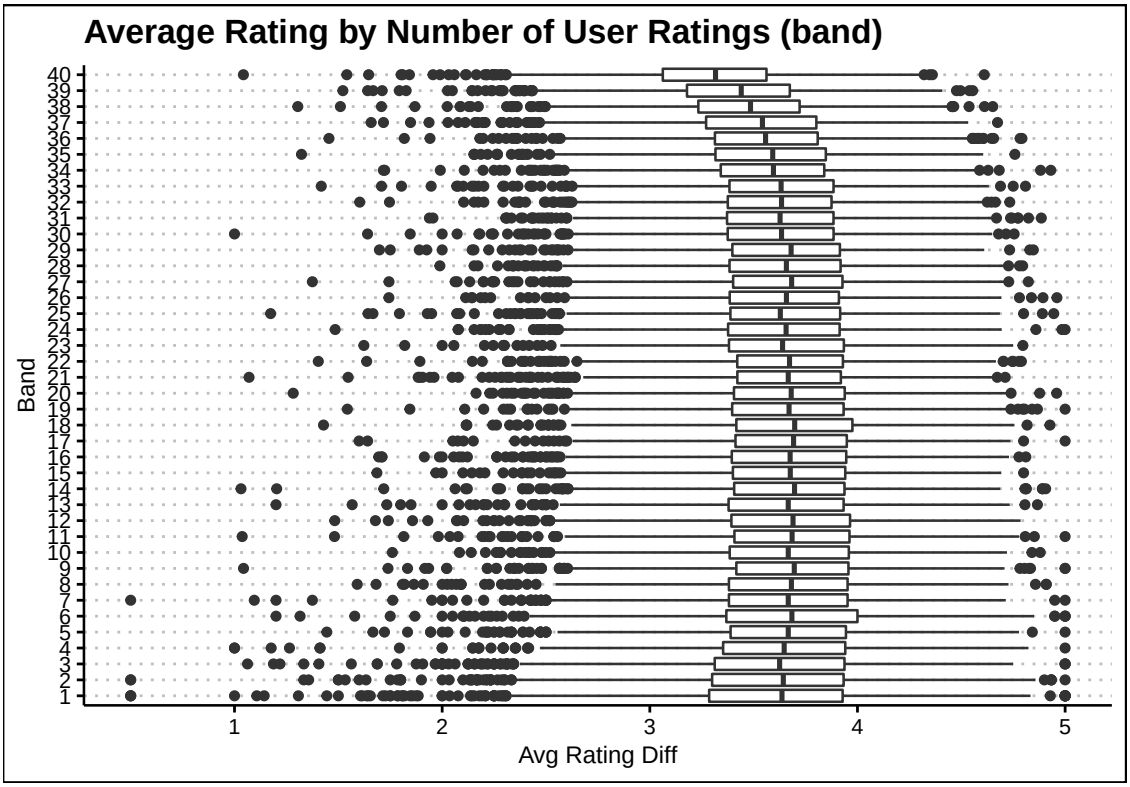


Is the individual user a factor in ratings?
The plot indicates that there is a difference in average rating of more than 1 between the lowest and highest rating users in this group. We can see that the user has an influence on ratings. The colour indicating the number of rating appears scattered randomly throughout users with various ratings so we do not expect the number of times a user rates to have a significant effect in this set of frequently rating users.

The following plot investigates any relationship between the rating and the number of times a user rates movies. Since the number of times a user provides a rating is varies widely, this has been divided into 40 bands using a balanced number in each band.



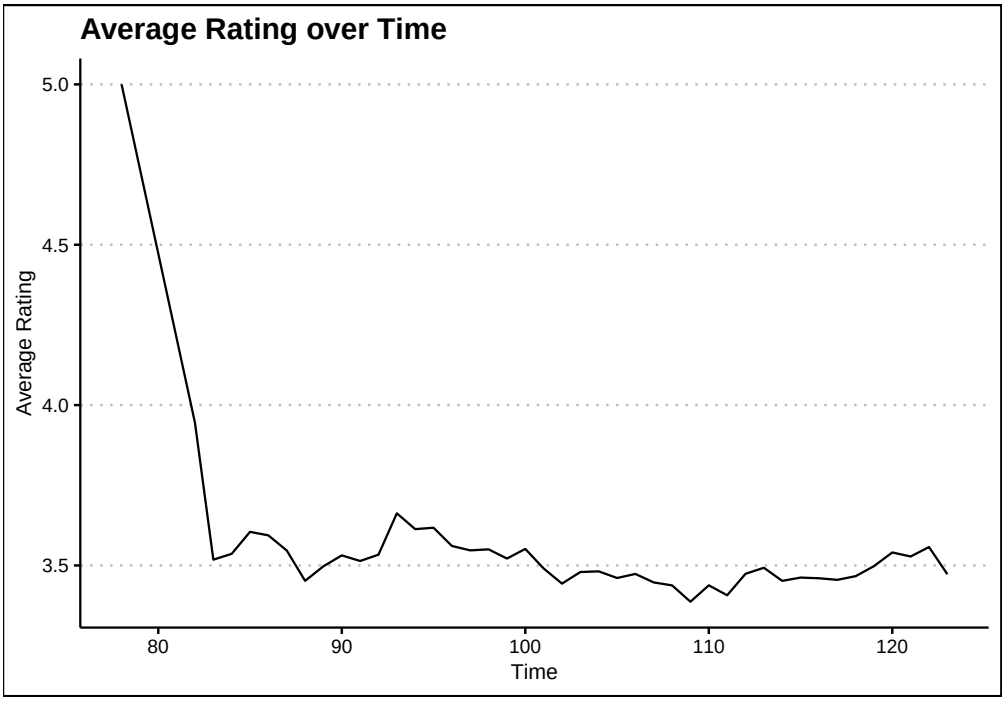
Is the number of ratings a user records a factor in ratings?
The plot indicates that there is no obvious difference in ratings in users that rate more frequently. This data however looks at the mean rating. in the following plot the range of ratings these bands take is investigated, and whether there are signficant outliers.

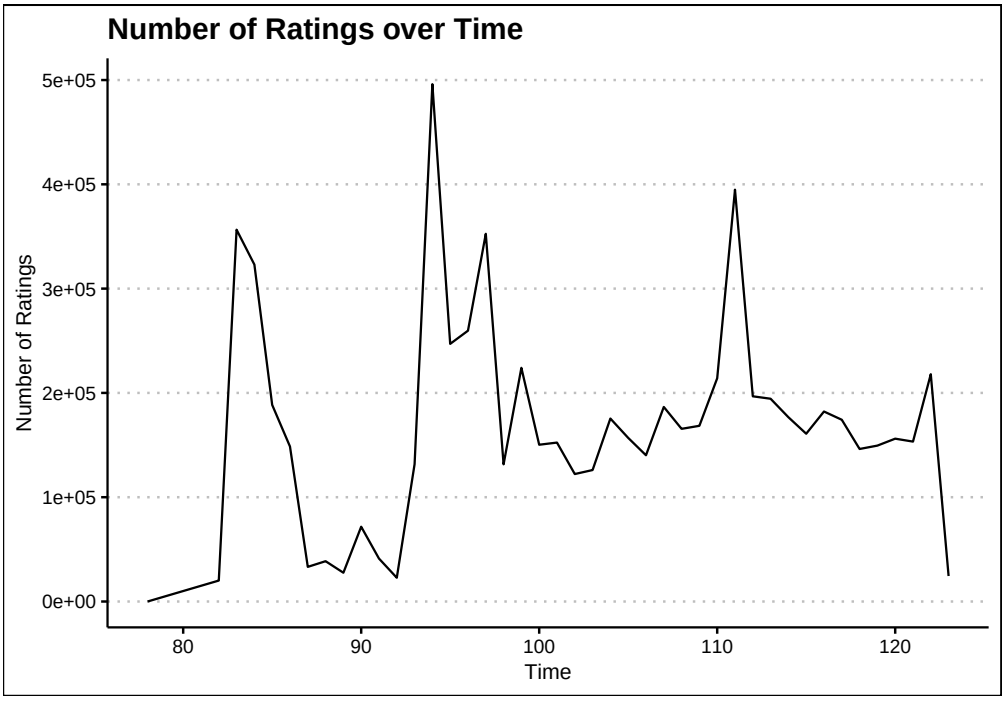


What is the range of ratings for users in a band, and are there signifiant outliers?
The plot clearly indicates that there is a slightly increased range of ratings in users that rate less often. There are significant outliers in all bands, with users of all rating levels demonstrating user specific rating levels. As a result of this finding the model will also need to regularize the predictions to reduce the variability of low rating users.

Time effect

The following plots examine whether time has an effect on the number or value of ratings.

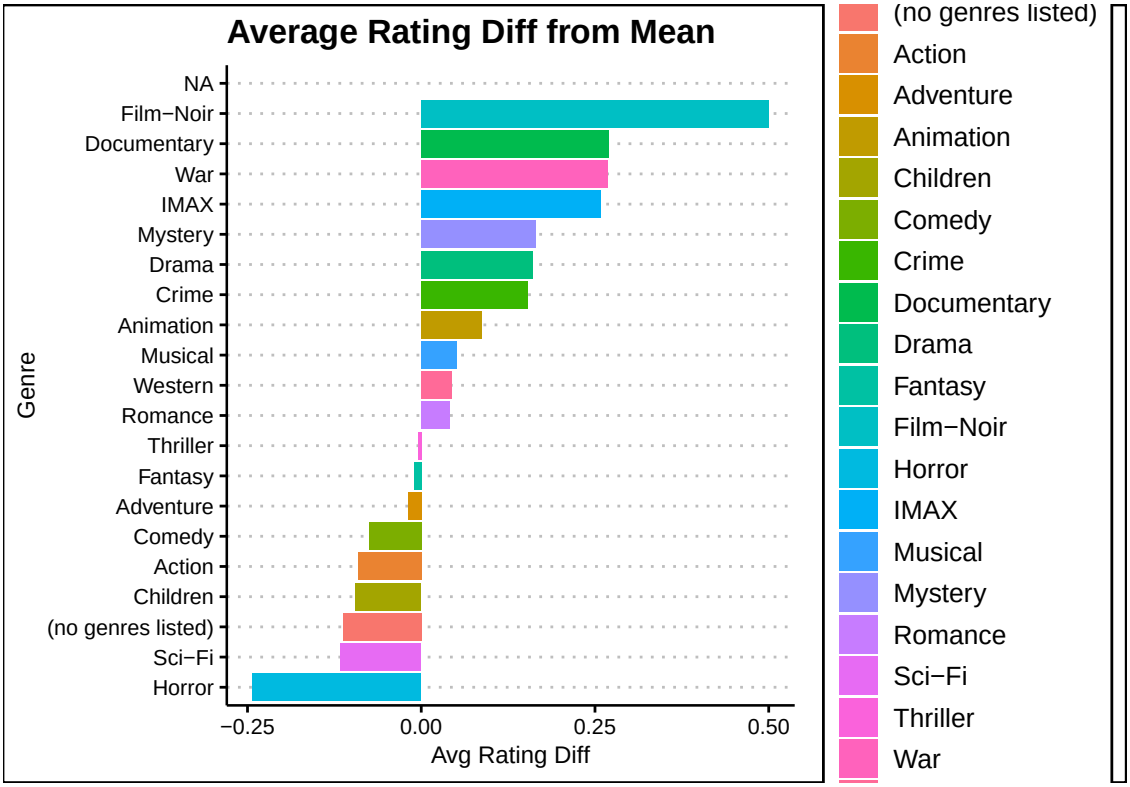




The average rating over time has remained fairly standard at around 3.5, with some variation over parts of the timeline. Initially higher ratings were indicated, but this matches a period of very low rating numbers as shown by the second plot. Time is included in the model, but it is not expected to provide a great deal of improvement in prediction.

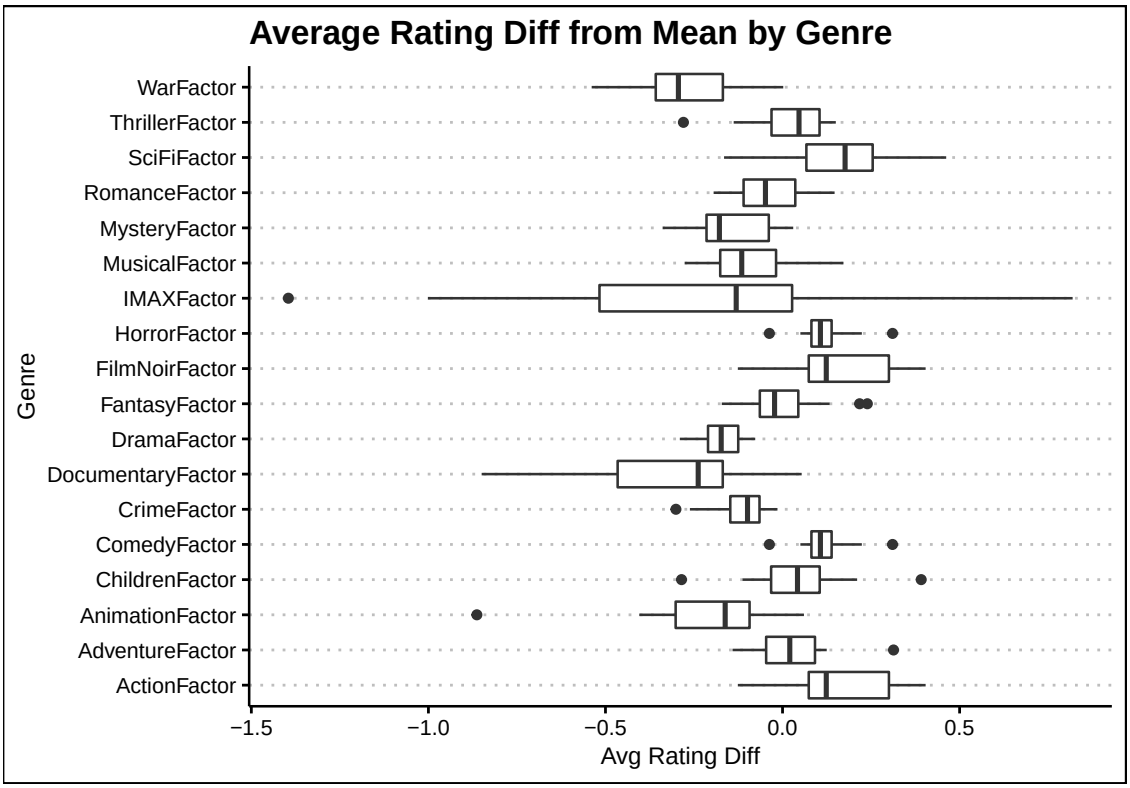
Genre effect

The first investigation is whether there is an overall genre effect. For example are Action movies more popular than Romances. The plot indicates the difference in ratings for a particular genre from the overall mean.



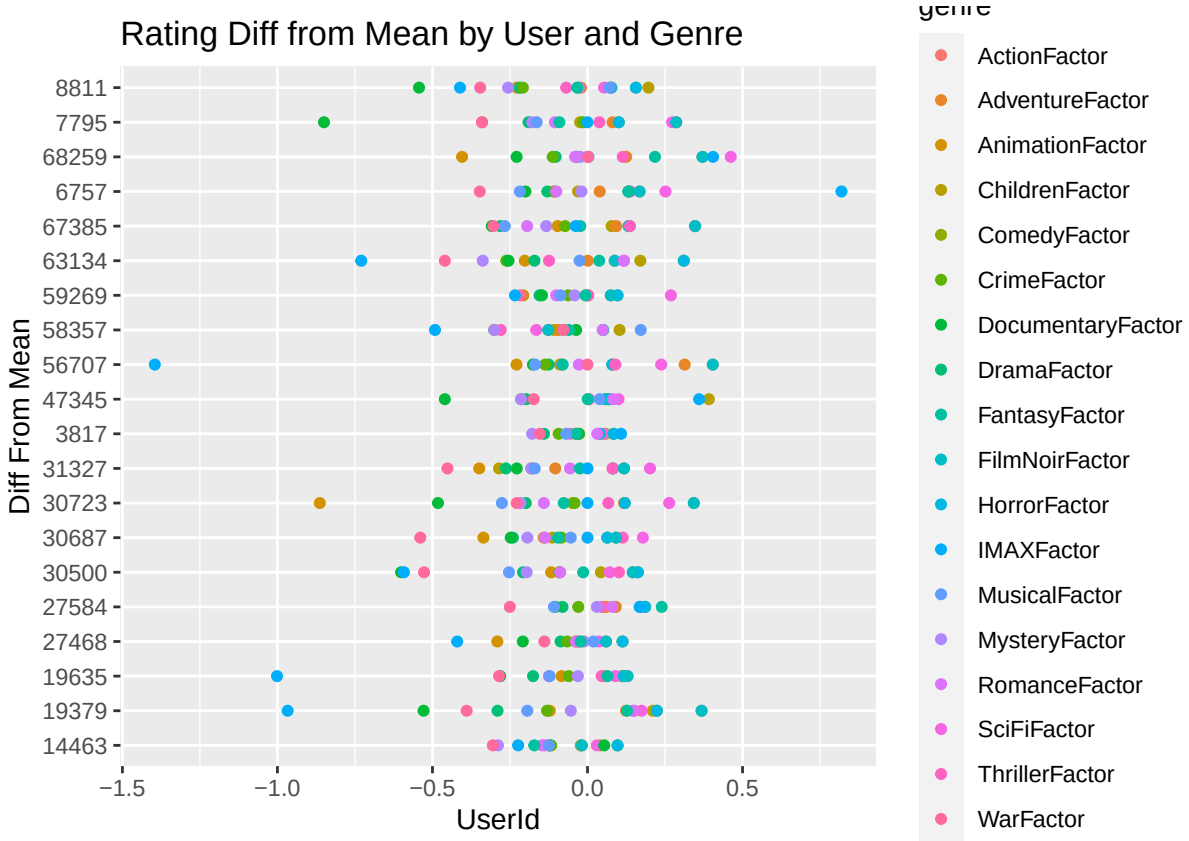
Some genres are clearly more/less popular than others.

The next plot looks at the range of difference in genre ratings from the overall mean rating. Some genres have a much wider range of ratings than others.



The following plot provides us with an indication of whether genre preferences are user specific. For each user the difference between the overall user mean and the user's mean rating for the genre is recorded. This provides a user factor for each possible genre (where the user has rated a movie of that genre previously).

Since there is no demographic data for the users, the best source of their preferences is likely to be the movie genres they have rated highly/poorly before. As a result, rather than applying a factor for genres overall the analysis investigates a user's preference for genres.



This plot provides evidence that genre preferences are very user specific. We can see, for example, that user 6757 rates horror movies very high, but user 56707 rates them very low.

The variation in colour patterns for each user is a clear indication that the model will be better able to predict a user's rating for a movie if it matches its genres to the user's genre preferences.

Model Creation and Evaluation

Model 1 - Movie and User

The initial model was based on the model used in Chapter 33 of the course materials. A penalised least squares approach for movie and user was used. In this case however the penalty term was obtained using cross validation with 7 samples from the train dataset.
The optimum value for the penalty term (lambda) was 4.75.
Predictions from the model were compared with the test set.
The Root Mean Squared Error (RMSE) for this approach was 0.8652421.

Model 2 - Movie, User and Time

The second model added a factor for the timestamp. As before the penalty term was obtained using cross validation with 7 samples from the train dataset.
Predictions from the model were compared with the test set.
The optimum value for the penalty term (lambda) was 5.
The Root Mean Squared Error (RMSE) for this apprach was .8651951.

Model 3 - Movie, User, Time and Genre

The third model added a factor for the user's genre preferences.

For each movie to be rated for the user, a user genre preference factor is included. For each genre that is applicable to a movie, the user factor is calculated as the difference between the user's overall mean and the user's mean for that genre. This factor is divided by the number of genres that apply to the movie.
The user genre factors are then summed for each genre applicable to the movie.

As an example, if Movie 4567 has 3 genres (Romance, Action and Comedy), the user genre factor is determined as being:
1/3 of the user's Romance factor + 1/3 of the user's Action factor + 1/3 of the user's Comedy factor

As before the penalty term was obtained using cross validation with 7 samples from the train dataset.
Predictions from the model were compared with the test set.
The optimum value for the penalty term (lambda) was 5.
The Root Mean Squared Error (RMSE) for this apprach was 0.8595488.

Final Results

Model 3 was selected as providing the best prediction. Predictions from the model were compared with the validation set.
The Root Mean Squared Error (RMSE) for this appr0ach was 0.859562.

It can be seen that while time had a minor effect, the genre data allowed for the most improvement in the base model.

Conclusion

This model uses all available data from the data set provided including Movie, User, Time and User/genre preference. The original Netflix Prize was an open competition which provided the entrants with only User, Movie, Grade, and Date of Grade.
Netflix' initial algorthim Cinematch scored an RMSE of 0.9514. The winning entrant, after 3 years scored an RMSE of 0.8567.

It is clear that this was quite an achievement. The RMSE of 0.8595488 obtained in this report had the advantage of a wider range of data, but less records, and significantly less time and resources. It could certainly be improved.

It should be noted however that some articles have questioned the validity of RMSE as a measure of a recommendation system. A 2009 TechnioCaliformia article (2) questioned the fact that the difference between a rating of 1 and 2 is measured the same as the difference beteen a rating of 4 and 5, however we would never recommend a movie with rating of 2.

It should also be noted that a number of law suits resulted from the Netflix competition, as even with the minimal information provided, some users were able to be matched to other systems and identified. (3)

Limitations

A very limited set of data was available for the users. User preferences could only be based on general movie and user ratings, and movies which they had previously rated, and which had similar genres.

The work was also limited by the hardware on which the models were run. Attempts to run more complex models (random forest and K nearest neighbours) resulted in vector memory errors and were abandoned.

Recommended Future Work

To improve the prediction a set of demograpghic data for users, for example age, gender, location, occupation, hobbies, could be used to identify users with similar details and to identify patterns in movie preference amongst users with similar details. This would be expected to improve the predictions, particularly for those users who have rated few movies.

It would also allow predictions to be made for users who have yet to rate any movies as long as the demographic data was avaiable.

Analysis of the data set on a hardware platform with more resources would allow additional modeling algorithms to be evaluated and combined for a final result. A wider range of techniques should be attempted, for example, matrix factorisation which was used in the Netflix solution. The results of these techniques could then be combined for a final result.

Technical Details

Saved data sets

The following data sets are provided with this report and allow the viewer to execute the report provided in .Rmd format: These data sets can be created by running the associated.R file, however the time taken for optimising the tuning parameters is significant and it is expected that the reviewer will load the provided data sets. edx.RData train_set.RData unique_genres.RData genre_data.RData top_20_train_pivot.RData user_genre_prefs.RData

---

References

(1) GroupLens, *MovieLens 10M Dataset*, accessed 16/12/2021, https://files.grouplens.org/datasets/movielens/ml-10m-README.html

(2) TechnCalifornia, Xavier Amatriain, *The Netflix Prize: Lessons Learned (2009)* accessed16/12/2021 http://technocalifornia.blogspot.com/2009/09/netflix-prize-lessons-learned.html

(3) Towards Data Science, Was Raham, *The Netfix Prize - How Even AI Leaders Can Trip up (2020)* accessed 16/12/2021 https://towardsdatascience.com/the-netflix-prize-how-even-ai-leaders-can-trip-up-5c1f38e95c9f