Movie Rating Prediction: EDX Capstone Course Part I

Megan Chandler

Introduction

Many organizations and websites now leverage Machine Learning capabilities to better anticipate and meet the needs and preferences of their customers. One such example is the film recomendation system used in video streaming sites such as Netflix, Amazon Prime Video, and YouTube. This project seeks to generate a Movie Recommendation System, utilizing Machine Learning processes, that takes into account rating differences between-Movies, -Raters, and -Genres to predict movie ratings.

This project used a subset of the data available from the GroupLens research lab dataset. This database includes ratings by more than 138,000 users for over 27,000 movies, resulting in a database with over 20 million ratings ranging from 1 to 5 stars. The dataset used for this project is a subset of the data from the database, resulting in a dataset with 10,000,054 ratings from 71,567 users and for 10,681 movies. These data were then split into two datasets, one consisting of 90% of the data to train the machine learning algorithm and the second, with the remaining 10% of the data, used to validate the Movie Prediction algorithm.

The final algorithm was developed after evaluating the impacts of various rating effects and determining the best-fitting model, evaluated in this project using RMSE. The final model leveraged an approach that accounted for rater, movie, and genre effects, resulting in a final RMSE of 0.8649. The following sections provide more detail on the project and results.

Method/Analysis

To train the algorithm, the initial training dataset (n = 9,000,055) was split further into a "Training Subset" (90% of original training dataset, n = 8,099,025) and "Validation Subset" (10% of original training dataset, n = 899,882) to be used for training purposes. Before splitting initial training dataset, the data were evaluated and further cleaned. Specifically, 36 raters were identified as having no variability in their ratings (e.g., a user rated 161 movies a 5). As the lack of variability of within-rater variability would not provide valuable information to inform the rating prediction algorithm, these 36 individuals were removed from the dataset, resulting in a final training dataset containing 8,998,899 ratings ($mean_{Rating} = 3.5$, $sd_{Rating} = 1.06$, $min_{Rating} = 0.5$, $max_{Rating} = 5$) from 69,842 unique raters for 10,668 movies.

In the training dataset, there is large variability in how many movies each rater has rated, with a minimum number of 10 ratings ($n_{Raters} = 1$) up to 6,615 ratings ($n_{Raters} = 1$); raters most rated 62 movies ($n_{Raters} = 406$). The movies represented a range of Genres, consisting of 796 Genre categories and one movie Pull My Daisy (1959) without a genre listed. Note that a Genre Catorgie is defined as the combination of genres assigned to a movie (e.g., Action, Action & Adventure), with each genre receiving at least 2 ratings (n = 8) and at most 733,172 ratings (n = 1, Drama). For training purposes, this dataset was then divided into the training and validation subset datasets. The training and validation datasets were adjusted to ensure that all movies and raters included in the validation dataset were also in the training dataset. See the table below for more information regarding the descriptives of the training subset data.

Rating Summary

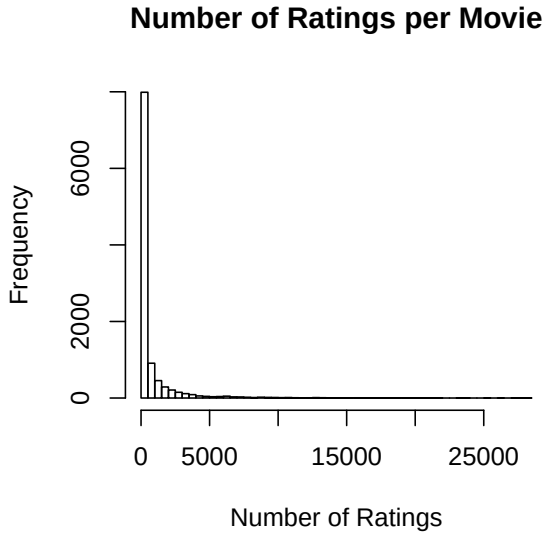
.	N	Mean	sd	Min	Max
Ratings	8099025	3.512382	1.060346	0.5	5

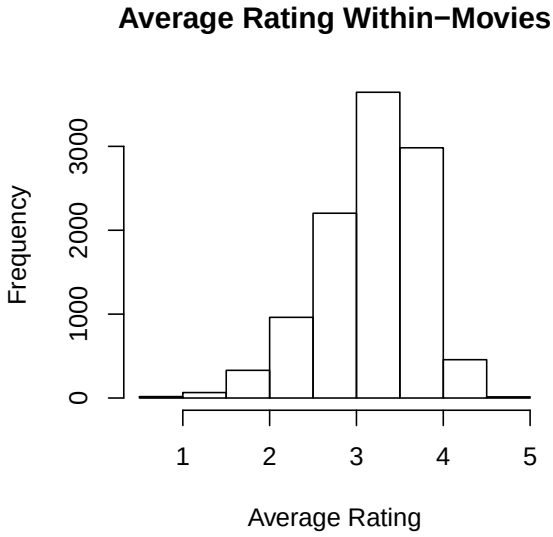
Initial exploration of the data showed that there was significant variability in the rating tendencies across raters, movies, and genres. These tendencies make it more difficult to predict a rating for a movie as the rating approach is not standardized across raters, movies or genres. To account for this, adjustments for an effect between movies, raters, and genres was made, as there were great differences in average ratings

within these factors. Specifically, when evaluating ratings within movies, wide variability in the number of ratings for each movie, as well as, average ratings within each movie was found (see plots below). This would make sense, as some movies included in the database are Blockbuster movies with millions of viewers and are therefore more likely to have a greater number of ratings compared to a small Independent film. For example, in evaluating the average ratings across raters, we see that some raters may generally rate movies higher than others and may show less variability in their rating (see plots below).

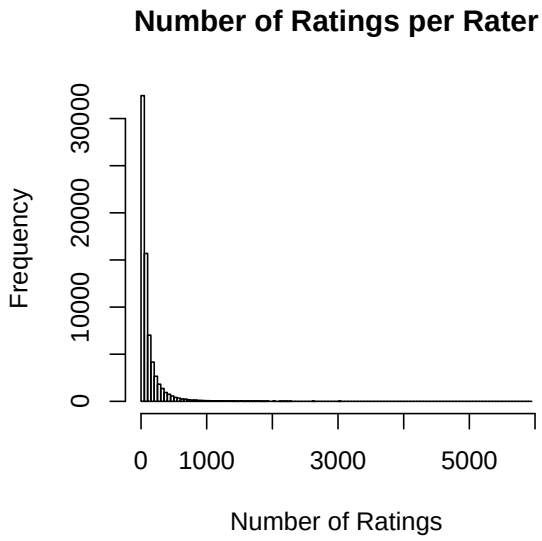
Within-Group Descriptives

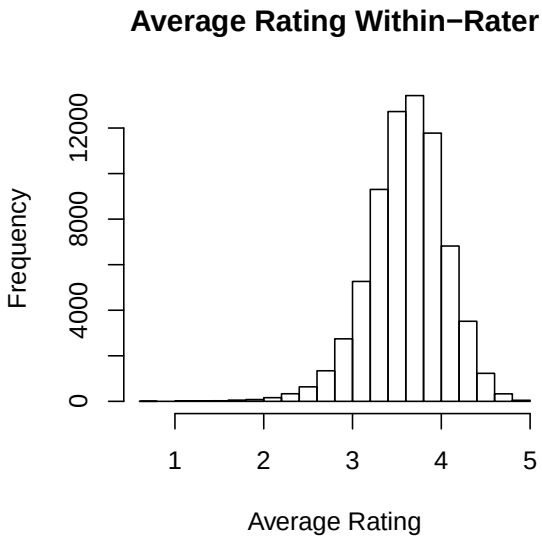
.	Median_N_Ratings	Min_N_Ratings	Max_N_Ratings	Within_Group_AverageRating_Min	Within_Group_AverageRating_Max
Raters	56	9	5935	0.7333333	5.000000
Movies	110	1	28257	0.5000000	5.000000
Genres	1303	1	659783	1.4574468	4.666667



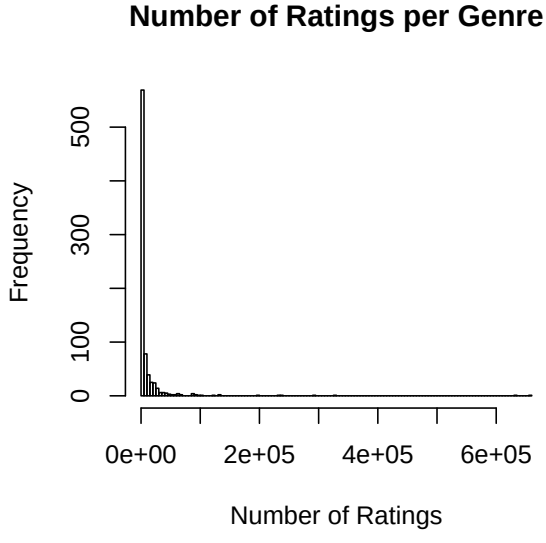


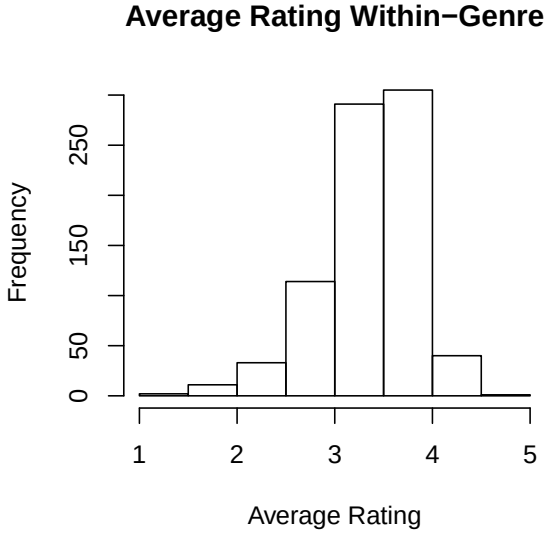
Similarly, variability in ratings across raters was observed. In other words, a high rating for Rater A (avg rating = 2, sd = 1) could be a 3 star rating, whereas this same rating of a 3 might be a low rating for Rater B (avg rating = 4, sd = 1) who generally rates movies higher on the scale. To be able to more confidently compare ratings across raters, account for this rater effect, and use this to predict ratings, all ratings were centered around the average rating for the dataset. As with betewen-rater variability, the next iteration of the model sought to adjust the predictions by taking into acount not only the Movie effect, but also the Rater effect.





Finally, as some genres may be more popular than others, it was expected the number of ratings and average ratings would have a high variability across genres. For example, movies within the Sitcom or Comedy genre will likely have more ratings as they are more popular compared to a "Film-Nior, Horror, Thriller". The data supported this as shown in the plots below. As the data suggested that genre did have an effect on the ratings, the effect of genre was taken into account in the final version of the model, in addition to the effects of the Movie and Rater.





Note that a date effect was also evaluated, and while there were some slight differences in rating tendencies across time, the date of the rating did not appear to have a significant effect.

When training the model, an iterative approach was used. The RMSE was evaluated for each model that was run was compared to a 'Naive Approach' as the starting point, which used the average rating (mean = 3.5) as the predicted value. With each iteration, the new model's RMSE was compared to the Naive

Model's RMSE and the RMSE of the previous models to see if there was an improvement. The final model, accounting for Movie, Rater, and Genre Effects, was then tested against the original validation dataset to evaluate the final RMSE.

Results

The table below demonstrates the improving RMSE with each iteration of the model, Naive Model ($RMSE_{init}$ = 1.06), Movie Effect Model ($RMSE_{init}$ = .9426), Rater & Movie Effect Model ($RMSE_{init}$ = .8644), and Movie, Rater, and Genre Effects Final Model ($RMSE_{init}$ = .8640). Note these models and the resulting RMSEs were evaluted against the training validation data set, that was a subset of the initial training dataset. Upon identifying the best performing model of those evaluated, the final model, Regularized Movie, Rater, and Genre Effects Model, was then used to predict ratings in the originally created Validation dataset, resulting in a final RMSE of 0.8649.

method	RMSE
Just the average	1.0591910
Movie Effect Model	0.9425787
Movie + User Effects Model	0.8643790
Movie + User + Genre Effects Model	0.8640464
Movie + User + Genre Effects Model- Final RMSE	0.8649469

Conclusion

The final model demonstrated a relatively strong prediction model for Movie Ratings, after accounting for effects from Raters, Movies, and Genres. This algorithm can be used to predict how raters may rate, or show preference for, various movies that they have not yet viewed or rated. Some limitations of the current model are that it does not take into account potential clustering, or group effects, among raters, and would be a potential next step in further developing this model. In other words, the model could evaluate the pattern of movie preference within raters and create groups, or clusters, of raters with similar preferences and rating tendencies. This could then be used to predict future ratings and make movie recommendations for a Rater based on what others in their cluster preferred.

Additionally, each rater, movie, and genre had a greatly varying number of ratings and the current model does not account for the number of ratings. In other words, a movie (rater/genre) with one rating carried the same weight in training the model as a movie (rater/genre) with 500 ratings. The account for this, the Movie, Rater, and Genre effects in future models can be regularized so that smaller samples within these groups are penalized and carry less weight in training the final model.