Movie Recommendation System

HarvardX Capstone

R Corrigan

2021 July 5

Overview

This report is a deliverable for the HavardX Capstone course, which is the ninth and final course of the HarvardX Data Science Professional Certificate Program. The challenge of the Capstone project is to create a movie selection system using the MovieLens dataset. The student is provided with a data set of movie ratings and is tasked with building an algorithm that predicts ratings that viewers give to movies. Once developed, the algorithm's utility is assessed by quantifying the closeness of predicted values to actual values on a separate validation set using the root mean squared error (RMSE) metric:

$$RMSE = \sqrt{\frac{1}{N}\sum(\hat{y}_{u,i} - y_{u,i})^2}$$

where N is the number of predictions, $\hat{y}_{u,i}$ is the predicted rating by user u of movie i and $y_{u,i}$ is the actual rating found in the validation set. Based on the grading rubric provided for the Capstone project, the RMSE component of the final grade achieves full marks if it is less than 0.86490. Thus the main technical criterion is to produce a model that predicts users' movie ratings with a RMSE of less than 0.86490.

The MovieLens 10M dataset, consisting of 10 million observations and available from http://grouplens.org, is used for the Capstone. HarvardX provided preliminary R code to refine the MovieLens dataset and partition it into two separate datasets: 1) an *edx* set, which is used to develop the prediction algorithm; 2) a *validation* set, which is used for the final test of the algorithm and calculation of the RMSE. The validation set is a partition of 10% percent of the MovieLens 10M dataset, and the remainder comprises the edx set. Thus, the edx set contains approximately 9 million observations (90% of the full dataset) for development of the algorithm. The edx data set is organized in a data frame. The first 10 rows are represented in the table below:

User ID	Movie ID	Rating	Time Stamp	Title	Genres
1	122	5	838985046	Boomerang (1992)	Comedy|Romance
1	185	5	838983525	Net, The (1995)	Action|Crime|Thriller
1	292	5	838983421	Outbreak (1995)	Action|Drama|Sci-Fi|Thriller
1	316	5	838983392	Stargate (1994)	Action|Adventure|Sci-Fi
1	329	5	838983392	Star Trek: Generations (1994)	Action|Adventure|Drama|Sci-Fi
1	355	5	838984474	Flintstones, The (1994)	Children|Comedy|Fantasy
1	356	5	838983653	Forrest Gump (1994)	Comedy|Drama|Romance|War
1	362	5	838984885	Jungle Book, The (1994)	Adventure|Children|Romance
1	364	5	838983707	Lion King, The (1994)	Adventure|Animation|Children|Drama|Musical
1	370	5	838984596	Naked Gun 33 1/3: The Final Insult (1994)	Action|Comedy

As shown, each observation includes a user, a movie, a rating, a time stamp, and applicable genres.

The dataset includes 10677 unique movies and 69878 unique users. The average rating is 3.51 with a standard deviation is 1.06.

Methods

To train a prediction algorithm, the edx dataset was further partitioned with 90% assigned to a train set and the other 10% retained as a test set. The rating prediction algorithm leverages the effects of features identified in the train set and uses regularization where appropriate. The predictors selected for the model are the movie, the user, the movie genre, and the time of rating. The potential prediction value of each of these features is discussed below.

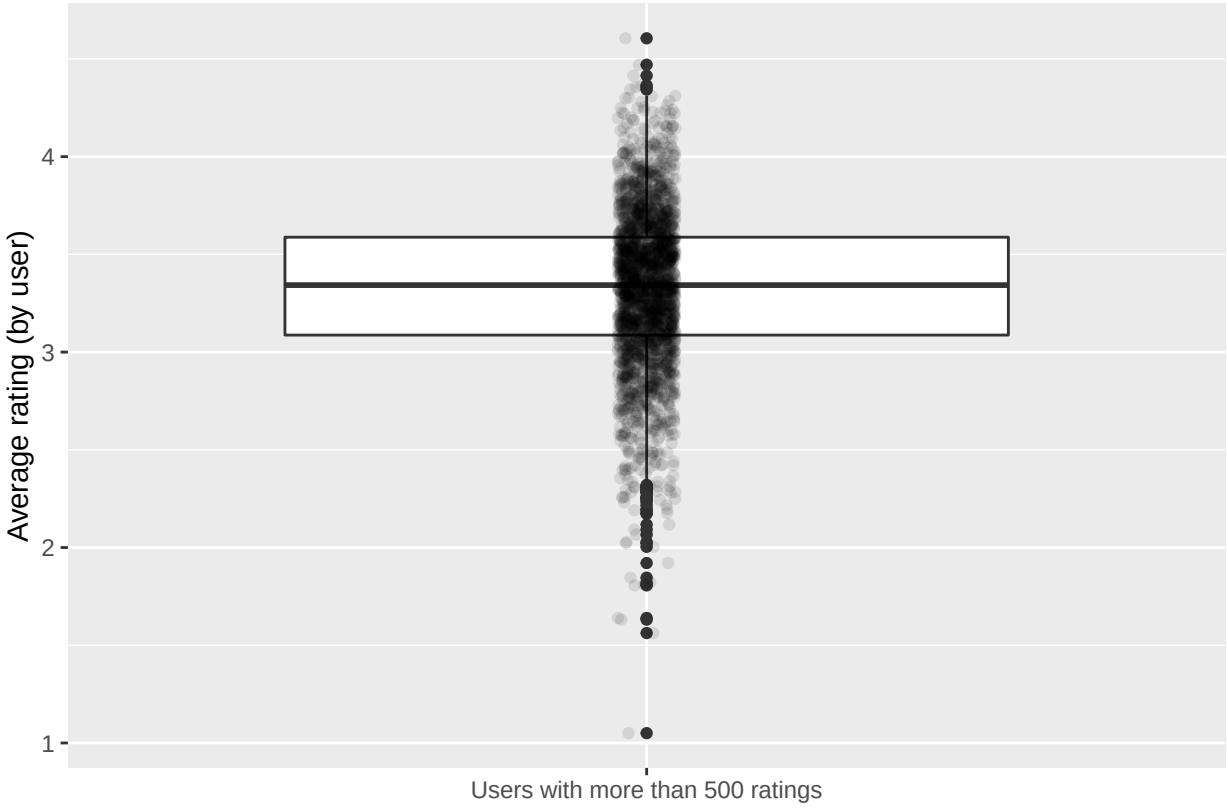
Predictors

Movie

Some movies are better than others and will garner higher ratings. In evidence, the three highest rated movies with more than 1000 ratings – *The Godfather*, *The Shawshank Redemption*, and *The Usual Suspects* – each have average ratings close to 4.5. Conversely, the bottom three movies with more than 1000 ratings have average ratings close to 1.7.

User

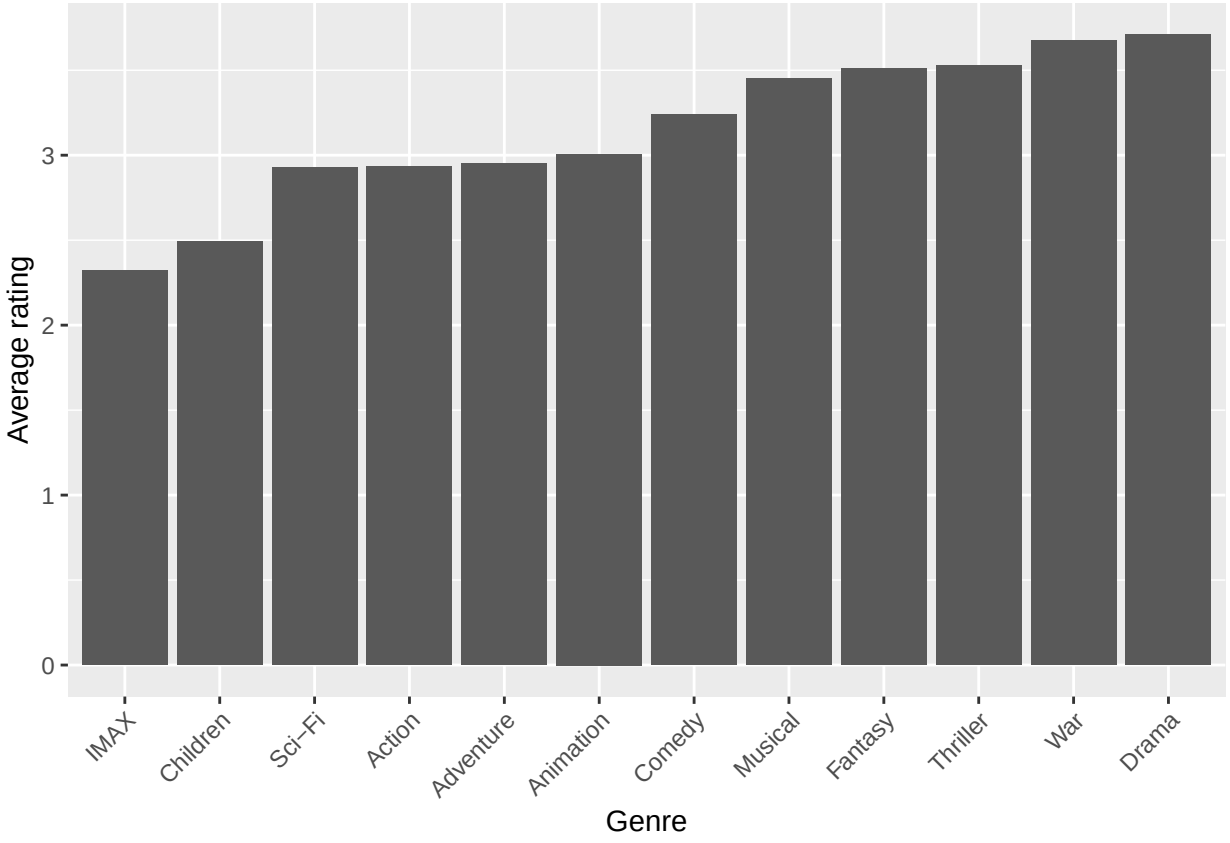
Users were found to have different biases, some having propensities to rate high and others to rate low. Even among users who have rated a large number of movies (i.e., more than 500), their average ratings vary substantially, as shown in the following plot.



The plot reveals a clear user effect. While it can be seen that the average of the users' average ratings is close to 3.3, some users are giving ratings of mostly 1 or 1.5 while others are giving ratings of mostly 4 or 4.5.

Genre

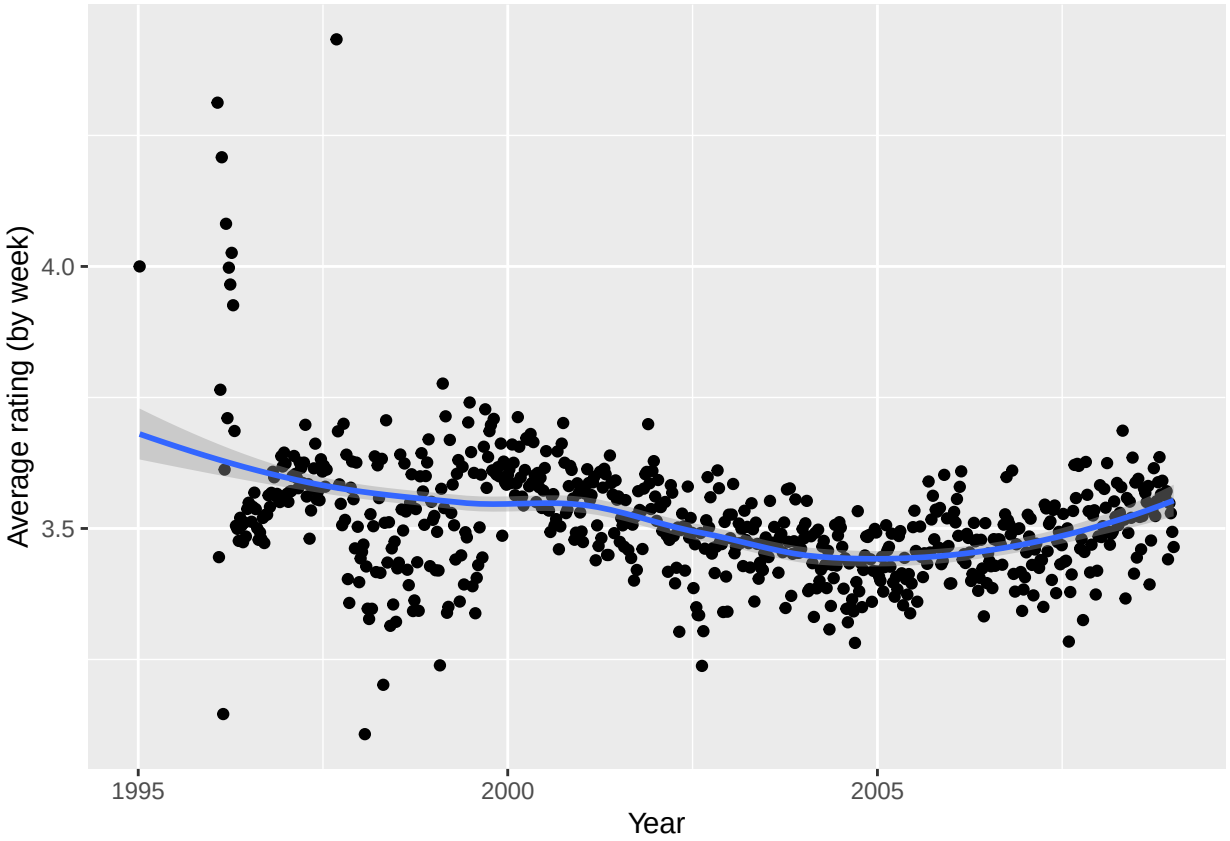
Most movies have multiple genre labels. For example, the movie *Outbreak* is labeled with four genres: action, drama, sci-fi, and thriller. In total, there are over 700 different combinations of genre labels for the movies in the edx set. For simplicity, the plot below includes only movies that are labeled with one genre to illustrate the effect of genre on rating.



Drama and war movies, for example, are rated an average of about 1.2 greater than IMAX and children movies. Thus, genre effects are a useful component of the model.

Time

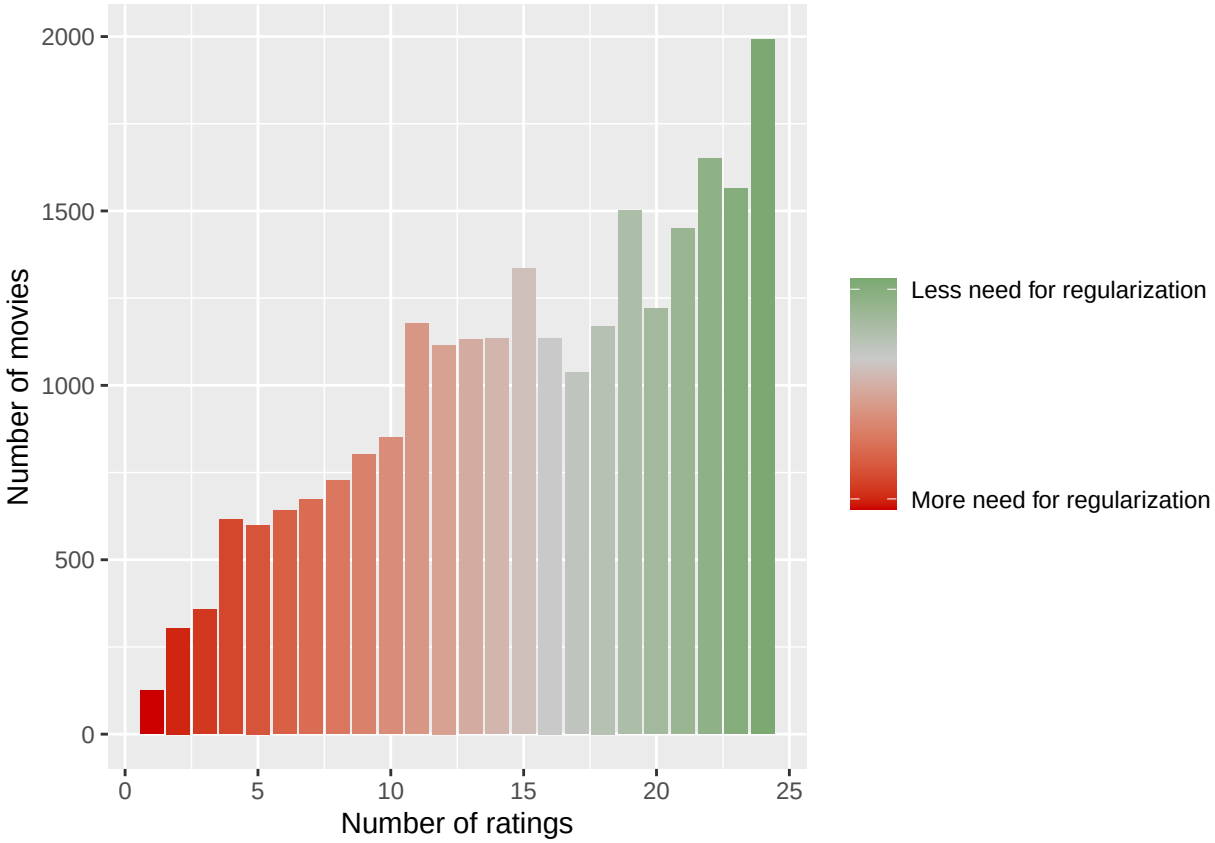
A scatter plot of ratings against time reveals a subtle time dependence of ratings.



We can see that movie ratings decline steadily until about 2005, after which they begin to ascend; modeling time effects can improve rating predictions. Based on visual assessment of the plot, estimating time effects at the resolution of quarter years (i.e. averaging ratings over bins of 3 months) is sensible.

Regularization

Regularization was employed to reduce the potential of large effects from small sample sizes. Some movies received few ratings and movie effects calculated for these movies are therefore prone to greater variance. It is desirable to assign less weight to these effects compared to effects derived from larger sample sizes. The plot below shows there are thousands of movies with 15 or less ratings.



There are also 51 genre combinations and 116 users associated with 15 or fewer ratings. With respect to time effects, stratifying the movie data by quarter-years results in no quarter with less than 170 ratings, except for 1995 for which there are only two ratings. Therefore, the two ratings from 1995 are excluded from analysis, and regularization is not necessary for the time effect term of the model.

Based on the abovementioned, regularization penalty parameters are incorporated for movie, user, and genre effects. The regularization parameters, $\boldsymbol{\lambda}$, were optimized by training the prediction model with the parameters set to integer values ranging from 0 to 20, and retaining the values that resulted in the lowest RMSE on the test set. Details of regularization calculations are provided in the following section on the model.

Model

Based on the evident rating effects from the features of movie, user, genre, and time, the following model for movie ratings is assumed:

$$Y_{m,u,g,t} = \mu + b_m + b_u + b_g + b_t + \epsilon_{m,u,g,t}$$

where $Y_{m,u,g,t}$ is the true rating for movie i by user u at time t; μ is the average rating for movie i; b_m, b_u, b_g, and b_t are the rating effects of movie, user, genre, and time, respectively; and $\epsilon_{m,u,g,t}$ is the combined error.

The movie effect, b_m was first calculated assuming a model that incorporates only the movie effect

$$Y_m = \mu + b_m + \epsilon_m$$

and estimating the movie effect as

$$\hat{b}_m = \frac{1}{n_m + \lambda_m} \sum (Y_m - \hat{\mu})$$

where λ_m is the regularization parameter for the movie effect and n_m is the number of distinct movies. Next, b_u was calculated assuming a model incorporating the movie and user effects

$$Y_{m,u} = \mu + b_m + b_u + \epsilon_{m,u}$$

and calculating the user effect as

$$\hat{b}_u = \frac{1}{n_u + \lambda_u} \sum (Y_m - \hat{\mu} - b_m).$$

This method was continued for estimating the remaining effects, b_g and b_t.

The model was then used to predict all movie ratings in the validation set, and the RMSE or these predictions was calculated to assess the utility of the prediction system.

Results

The optimal values for the regularization penalties, λ_m, λ_u, and λ_g, were found to be 4, 4, and 17, respectively. These were optimal in the sense that they provided the smallest RMSE on the test set, which was 0.86372. With these parameter values incorporated in the model, and with the effects calculated for all movies, users, genres, and times (quarters), movie ratings were predicted for the validation set. The predicted ratings resulted in a RMSE of 0.86478, thus meeting the highest technical criterion of a RMSE of less than 0.86490.

Conclusion

A movie recommendation system was developed using the Movie Lens 10M data set. The dataset was partitioned so that 90% percent used for training and testing of the prediction algorithm, while the other 10% was sequestered as the validation set for assessing the algorithm's performance. The developed algorithm predicts ratings of movies by using a simple model that incorporates estimated effects from four features in the data set: movie, user, genre, and time of rating. Regularization was applied in the calculation of the movie, user, and genre effects due to the small number of observations for some instances of these features. The final test of the algorithm on the validation set produced a RMSE of 0.86478 for the predicted ratings.