MovieLens project

Ben Neely

February 10, 2019

OVERVIEW

Data scientists are often tasked with navigating large data sets to develop predictive models. In the current project, a database consisting of 10,000,054 movie ratings applied to 10,681 movies by 71,567 users was analyzed. The purpose of this exercise was to develop and evaluate a predictive model consisting of various factors on movie ratings. Specifically, the objective was to develop a predictive algorithm from a training data set (i.e., 10% of total ratings) that when applied to the validation data set (90% of total ratings) produced a root mean squared error (RMSE) value <= 0.8775. A suite of models containing specific movie, user, release year, review year, and genre effects were evaluated and predictive ability for each was measured by RMSE. When all predictive factors were included in the model, RMSE = 0.8542. Despite meeting the RMSE objective, accurate ratings were only predicted in 36% of occurrences when the model was applied to the validation data set. This is attributed to actual ratings being in increments of 0.5 while predicted ratings could be any real value between 0.0 and 5.0. The RMSE value meeting the objective and difficulities associated with accurately predicting rating suggests that the algorithm presented herein was a suitable predictor of movie ratings.

METHODS

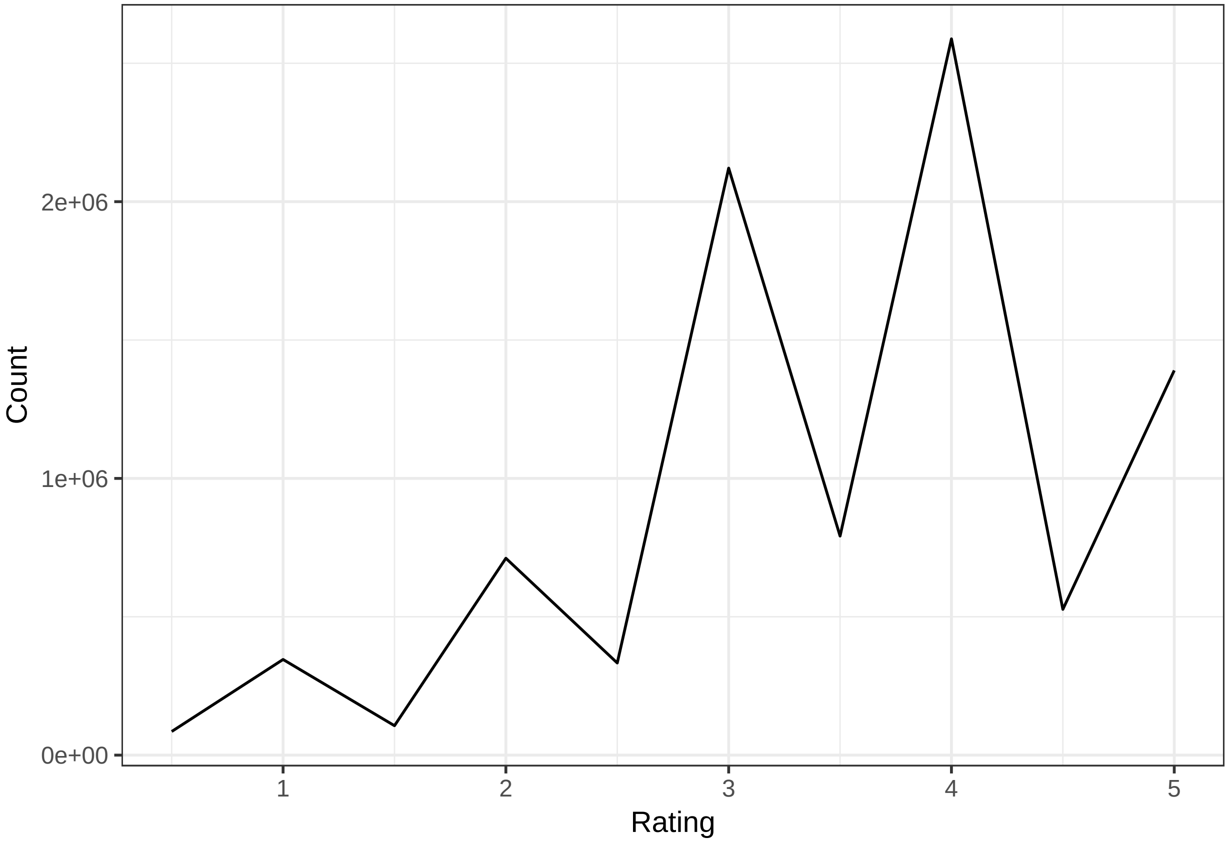
Data used for this exercise were contained in the MovieLens 10M dataset developed and maintained by the University of Minnesota. Details about the data set can be found in the README file associated with the download data. The original data set consisting of 10,000,054 movie ratings applied to 10,681 movies by 71,567 users was split into a training data set (10% of ratings) and a validation data set (90% of ratings). Before algorithm development began, data were cleaned in three ways. First, a function was created to extract release year from the movie title. Second, genres associated with each movie rating were separated. For example, a single rating on a movie with original genre as action, crime, and thriller was separated into three separate ratings that shared the same information except only one genre was included. This allowed evaluation of individual genres as predictive factors. Third, the timestamp was converted to an understandable review year.

```
substrRight <- function(x, n){
  substr(x, nchar(x)-n+1, nchar(x))}

edx1=edx%>%mutate(year=substrRight(title,5))
edx2=edx1%>%mutate(year=substr(year,1,4),
                   year=as.integer(year))

edx3=edx2%>%separate_rows(genres,convert=T)

edx4=edx3%>%
  mutate(date=as_datetime(timestamp))

edx4$rev_year=year(edx4$date)
```

Following data cleansing, a function was developed to calculate RMSE for each model.

```
RMSE=function(true_ratings,predicted_ratings){
  sqrt(mean((true_ratings-predicted_ratings)^2))}
```

The intital predictive model simply used mean of all ratings as the lone predictive variable. Movie, user, release year, review year, and genre effects were iteratively added and RMSE was calculated for each model to measure fit. The best fit model was applied to the validation data set to predict ratings. Predicted ratings were rounded to the nearest integer because ratings were more likely to be integers than in 0.5 increments. Accuracy was determined as the percentage of predicted ratings that matched actual ratings.



RESULTS

The model containing movie, user, release year, review year, and genre effects resulted in the lowest RMSE (0.8542). However, RMSE would have met the objective if only movie and user effects were considered. Release year, review year, and genre added little predictive capability once movie and user effects were considered. Assuming the objective of this exercise is to minimize RMSE, the model containing all explanatory variables should be used to predict movie ratings.

Model	RMSE
Average rating	1.0547271
Movie Effect	0.9416321
Movie + User Effects	0.8547183
Movie + User + Release Year Effects	0.8543808
Movie + User + Release year + Review year Effects	0.8542917
Movie + User + Release year + Review year + Genre Effects	0.8541967

CONCLUSIONS

Data science skills including data organization, data cleansing, algorithm development, and data visualization were used in this exercise to predict movie ratings from a large data set. Root mean squared error values decreased with inclusion of more explanatory variables. However, simply using movie and user effects to predict movie ratings would have been sufficient in this instance. Despite the acceptable RMSE value produced by the model that incorporated movie, user, release year, review year, and genre effects, accuracy of the model was only 36% when applied to the validation data set. The reason for this is that the model predicted continuous values for movie ratings while actual movie ratings were discrete (0.5 increments). This led to inherent errors in rounding that lowered model accuracy. In instances where objectives are similar to those in this example, data scientists should examine many sources of variation in the dependent variable and incorporate them strategically into predictive models.