Predicting Movie Ratings Report

Dom Pizzano

12/30/2020

Predicting Average Rating of Movie with Number of Ratings per Movie and the Release Year

Overview:

Calculating the average movie ratings using the number of ratings and the year released of each movie. Using data from https://grouplens.org/datasets/movielens/10m/ for the ratings. The data comes in the format of total ratings and has 9000055 total ratings for all movies in the data set. By grouping and running analysis on the data, it will become summarized for each movie, then using a multiple factor linear regression model, predict the average rating for each movie, then compare it to the actual average value of the movie using the RMSE (root mean square error) to determine how accurate our prediction model is. Then repeating the steps to see if running a linear regression model with just the rating count will produce a more accurate prediction model than the mutiple variable (count + year) linear regression model.

First to load in the data as used in the EDx course to get our edx dataset with the total ratings.

There are 9000055 rows/ratings in the initial dataset.

Methods/Analysis:

Next, to process the nine million rating records from the dataset created, there are a few steps to clean the data and manipulate it to get the summarized data format needed for the analysis.

1.) Grouping the data by movidId
2.) Mutating/adding data columns for our new summarized data
3.) Calculating the total number of ratings for each movie
4.) Calculating the sum of total ratings for each movie
5.) Calculating the average rating for each movie by dividing the sum of total ratings by the total number of ratings for each movie
6.) Extracting the year of the movie from the title using a string split and string remove all parenthesizes
7.) Filtering out the movies that have less than 9 ratings because their averages are very skewed because of the low number of ratings
8.) Only collecting the unique values of the movieId's
9.) Arranging by average rating descending (to see which movies have the highest average rating)

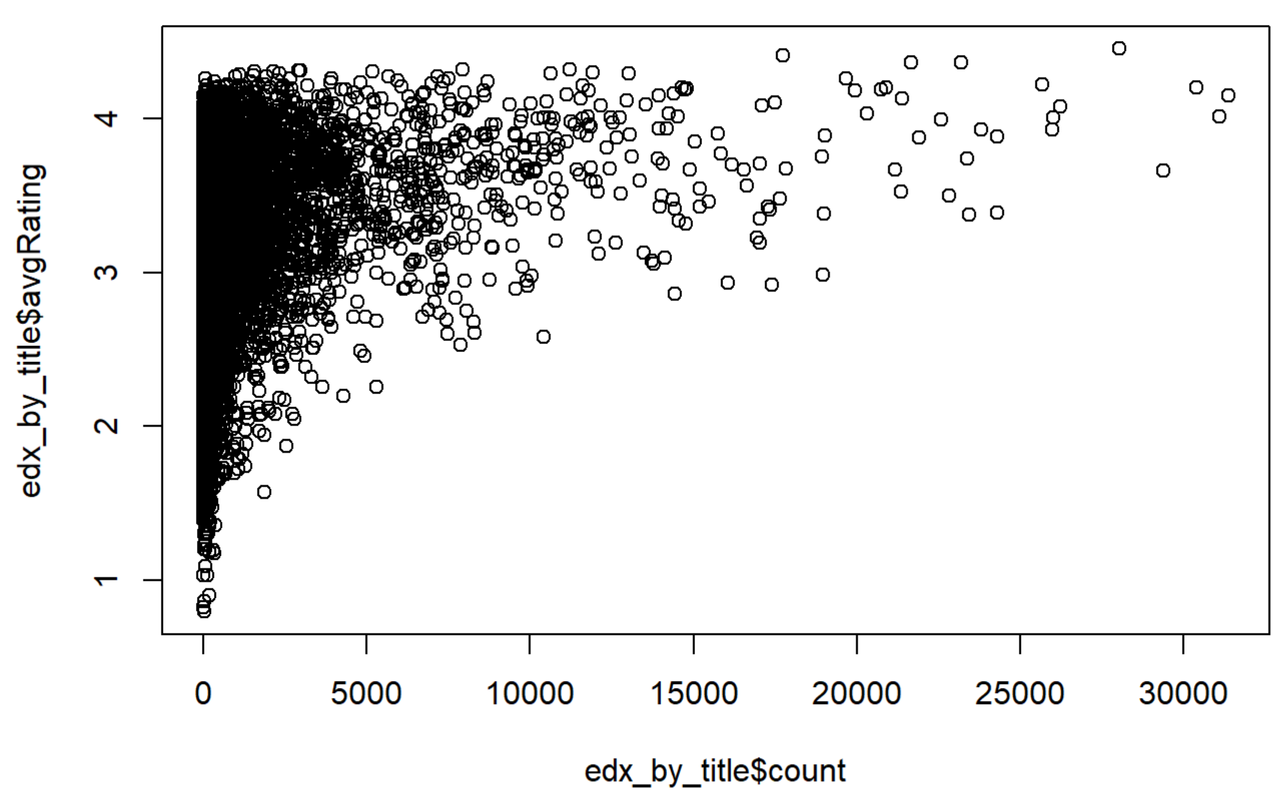
Results of Data Pre-Processing:

```
edx_by_title <- edx %>% group_by(movieId) %>%
  mutate(RatingSum = sum(rating),
  count=n(),
  avgRating=RatingSum/count,
  year= strtoi(str_remove_all(str_sub(title,-5,-1),"[()]"))) %>%
  summarise(title,avgRating,year,count) %>%
  filter(count>=9) %>%
  unique(.) %>%
  arrange(-avgRating)
head(edx_by_title,10)
```

movieId <dbl>	title <chr>	avgRating <dbl>	year <int>	count <int>
318	Shawshank Redemption, The (1994)	4.455131	1994	28015
858	Godfather, The (1972)	4.415366	1972	17747
50	Usual Suspects, The (1995)	4.365854	1995	21648
527	Schindler's List (1993)	4.363493	1993	23193
912	Casablanca (1942)	4.320424	1942	11232
904	Rear Window (1954)	4.318652	1954	7935
922	Sunset Blvd. (a.k.a. Sunset Boulevard) (1950)	4.315880	1950	2922
1212	Third Man, The (1949)	4.311426	1949	2967
3435	Double Indemnity (1944)	4.310817	1944	2154
1178	Paths of Glory (1957)	4.308721	1957	1571

1-10 of 10 rows

To view how the data looks, plot the relation of ratings count and the average rating of each movie. To see if count could be a good predictor of average rating, create linear models and run an RMSE, also to see how year released affects the prediction, that will be a factor in one of the models.



Create the linear regression models for the data frame edx_by_title column average rating using rating count and year released for each movie.

```
> lmmov <- lm(avgRating ~ count + year, data=edx_by_title)
```

Creating a linear regression for average rating just based on the rating count for each movie.

```
> lmmov2 <- lm(avgRating ~ count, data=edx_by_title)
```

Predicting the average ratings of each movie use the predict function with the linear models created above.

```
> newdata <- data.frame(count=edx_by_title$count,year=edx_by_title$year)
> predicted_Ratings <- predict(lmmov,newdata)
> newdata2 <- data.frame(count=edx_by_title$count)
> predicted_Ratings2 <- predict(lmmov2,newdata2)
```

Results:

Here is a preview of the data with the average rating compared to the predicted average rating for the first linear model created (lmmov).

```
> edx_by_title$prediction <- predicted_Ratings
> edx_by_title %>% select(title,avgRating,prediction) %>% head(n=5)
```

movieId <dbl>	title <chr>	avgRating <dbl>	prediction <dbl>
318	Shawshank Redemption, The (1994)	4.455131	4.684183
858	Godfather, The (1972)	4.415366	4.260646
50	Usual Suspects, The (1995)	4.365854	4.317618
527	Schindler's List (1993)	4.363493	4.419042
912	Casablanca (1942)	4.320424	4.105785

5 rows

Here is the statistical summary that shows the significance of each linear model. The p values here show how significant the variable is in determining the predicted average rating.

```
> summary(lmmov)
```

```
Call:
lm(formula = avgRating ~ count + year, data = edx_by_title)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.24005 -0.31322  0.06795  0.38900  1.21117 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.726e+01  5.698e-01   30.29   <2e-16 ***
count        5.646e-05  2.274e-06   24.83   <2e-16 ***
year        -7.099e-03  2.868e-04  -24.75   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.521 on 9709 degrees of freedom
Multiple R-squared:  0.1061,	Adjusted R-squared:  0.1059 
F-statistic: 576.4 on 2 and 9709 DF,  p-value: < 2.2e-16
```

```
> summary(lmmov2)
```

```
Call:
lm(formula = avgRating ~ count, data = edx_by_title)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.36299 -0.33402  0.07321  0.40071  1.10510 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 3.155e+00  5.866e-03  537.82   <2e-16 ***
count       5.272e-05  2.339e-06   22.54   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5372 on 9710 degrees of freedom
Multiple R-squared:  0.04973,	Adjusted R-squared:  0.04963 
F-statistic: 508.1 on 1 and 9710 DF,  p-value: < 2.2e-16
```

Comparing the RMSE of the linear models to see which one was better at predicting the average rating. Do this by comparing the actual average ratings to the predicted average ratings.

```
> rmse1 <- rmse(edx_by_title$avgRating,predicted_Ratings)
> 
> rmse2 <- rmse(edx_by_title$avgRating,predicted_Ratings2)
```

Linear model (count+year variables) RMSE: 0.520949
vs
Linear model (count variable) RMSE: 0.5371316

Conclusion:

The conclusion that can be drawn from the RMSE's above is that it is more accurate to have a multiple variable linear regression model with count and year variables to predict the average movie rating than it is to just use count to predict the average movie rating. Both variables year and the count had p values that showed it significantly could predict the average rating of the movie. Both linear models RMSEs were less than .8649, a goal for the assignment.