MovieLens Capstone

Greg L

3/1/2020

Introduction

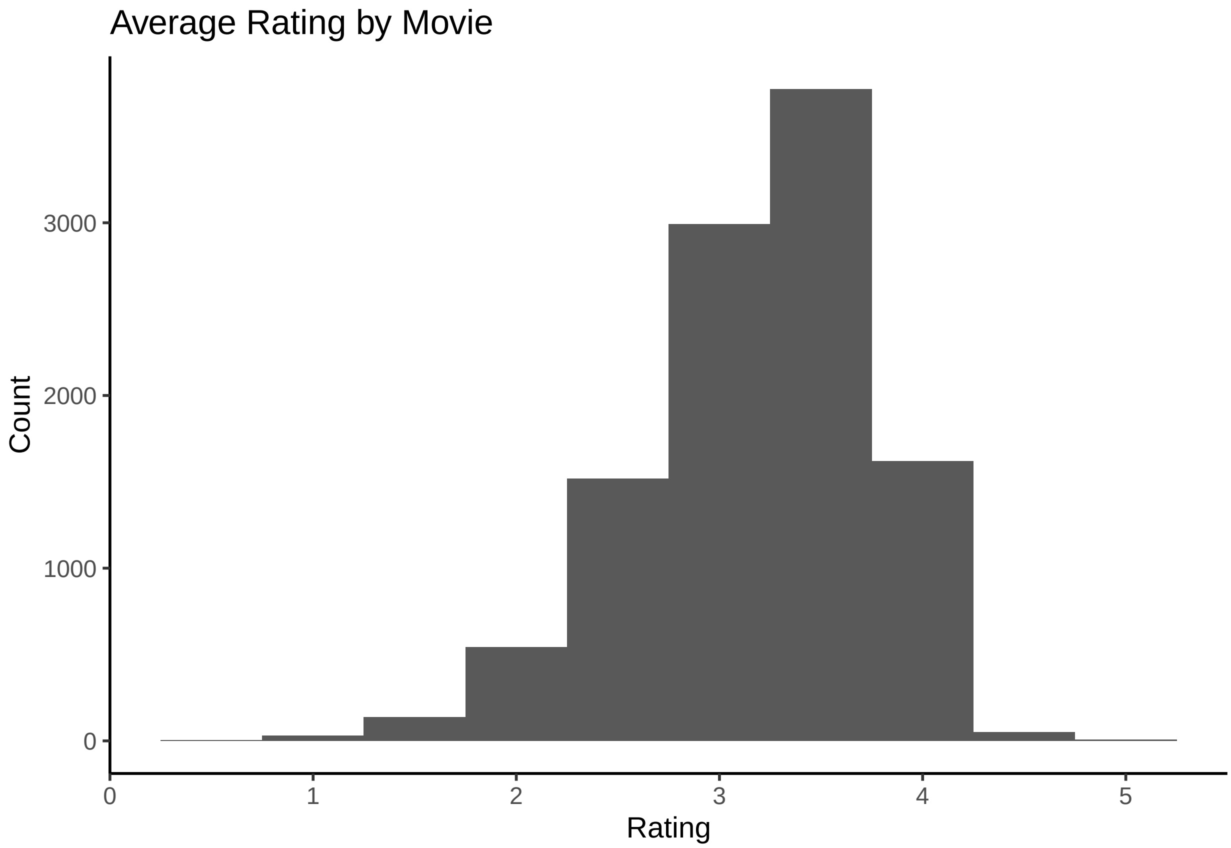
This project is designed to create a prediction model for movie ratings based on the MovieLens data set. The full 10M row set will be used and split into a training set and validation set. The validation set is 10% of the entire data set. The preview below shows the metadata associated with each movie rating in the data. Data models were created and tested to find the best combination of factors that minimize RMSE (Root Mean Square Error).
Files were downloaded from: *http://files.grouplens.org/datasets/movielens/ml-10m.zip*

```
head(edx)
```

```
##   userId movieId rating timestamp                        title
## 1      1     122      5      1996             Boomerang (1992)
## 2      1     185      5      1996              Net, The (1995)
## 4      1     292      5      1996              Outbreak (1995)
## 5      1     316      5      1996              Stargate (1994)
## 6      1     329      5      1996 Star Trek: Generations (1994)
## 7      1     355      5      1996      Flintstones, The (1994)
##                           genres
## 1                 Comedy|Romance
## 2          Action|Crime|Thriller
## 4  Action|Drama|Sci-Fi|Thriller
## 5        Action|Adventure|Sci-Fi
## 6 Action|Adventure|Drama|Sci-Fi
## 7       Children|Comedy|Fantasy
```

Methods/Analysis

The initial test is to see whether the simple averaging of movie ratings acts as a reliable predictor of ratings.



```
# A tibble: 1 x 2
  method             RMSE
  <chr>             <dbl>
1 Just the average  1.06
```

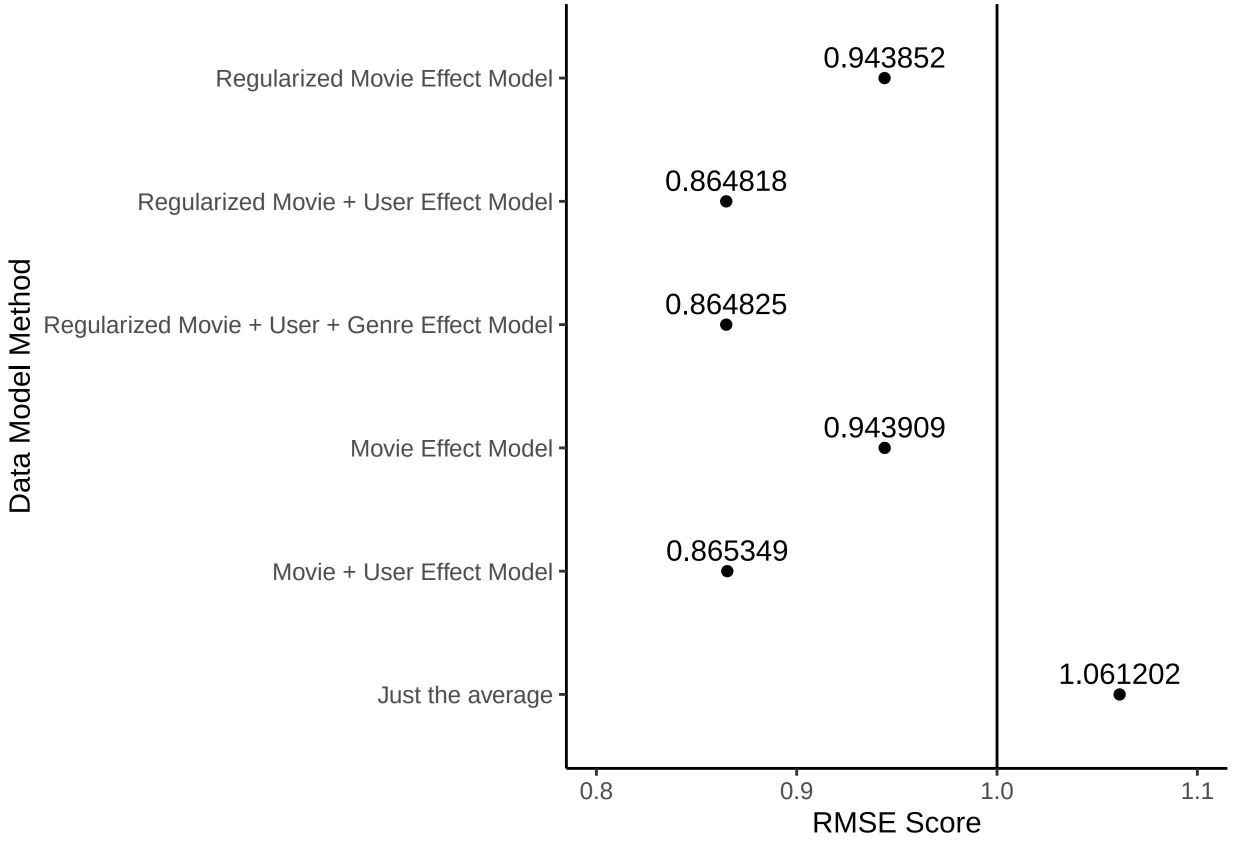
Since the RMSE is above 1.0 the model needs to be improved.

The first two models evaluate the predictive power of a) movies themselves or b) both movies and users on ratings.

```
# A tibble: 3 x 2
  method                      RMSE
  <chr>                      <dbl>
1 Just the average           1.06
2 Movie Effect Model         0.944
3 Movie + User Effect Model 0.865
```

An attempt was made to regularize the data sets and then evaluate predictive power of several options:

- Regularized Movie Effect Model
- Regularized Movie + User Effect Model
- Regularized Movie + User + Genre Effect Model



Final Results

Several models were designed and tested against the validation set. The result is below The best RMSE score is:

```
[1] "Table of RMSEs"
```

method	RMSE
Just the average	1.061202
Movie Effect Model	0.943909
Movie + User Effect Model	0.865349
Regularized Movie Effect Model	0.943852
Regularized Movie + User Effect Model	0.864818
Regularized Movie + User + Genre Effect Model	0.864825

```
[1] "The best model is the: Regularized Movie + User Effect Model"
```

```
[1] "It has an RMSE score of: 0.864818"
```

As we can see, adding parameters to the model increases the predictive capabilities, but up to a point at which the model begins to over-fit and the predicted ratings begin to vary more from the actual rating.

Conclusion

Movie ratings can be predicted by analyzing trends in user behavior, among other factors. Prediction models should be tested first to find the proper categories to use that would minimize bias and overfitting of models. In the MovieLens model, incorporating Genres tended to increase the errors of predictions instead of improving. Future work can include larger data sets and testing of seasonality or years in which ratings are made for films to see if that influences a movie's rating.