Movie Recommendation

Gaurav

1 March 2019

GitHub

Introduction

This project involves the prediction of movie ratings for different users.
The dataset is given in the HarvardX Data Science Capstone Project.

The data is partitioned into a training set with ~9million entries and 100k validation entries.

Each entry in the dataset is a rating given by a user. It contains the movie title, the userId, movie Id,the genres in that movie and rating given by user.

The problem statement is effectively a table with movieId on x axis, userId on y axis and ratings in the cells.
The table is sparesely filled and the challenge is to fill the blanks. The table is also large.

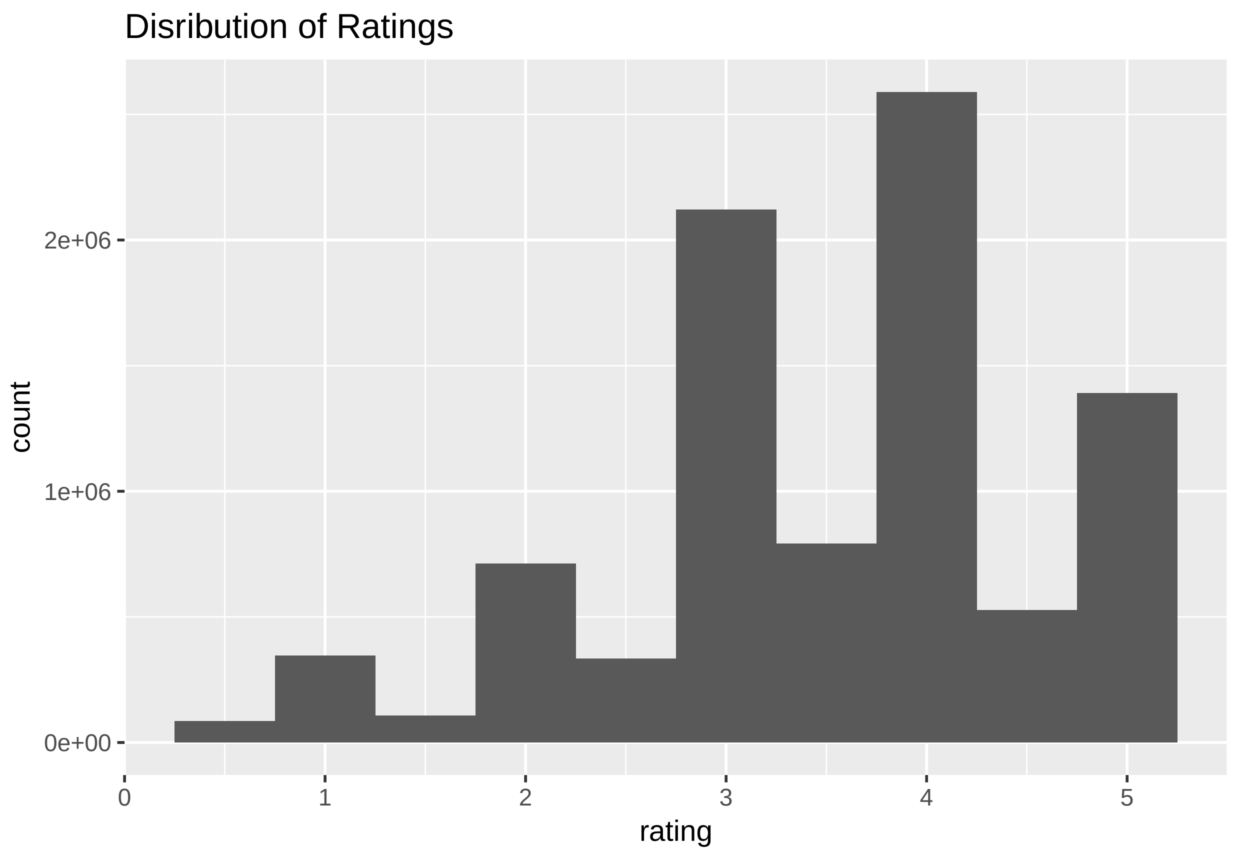
My solution to the project is a model based approach.

> Rating = Mu + b_i + b_u + b_g
> where
> Mu = Average Rating for all Movies
> b_i = Movie Effect
> b_u = User Effect(Overall)
> b_g = User Genre Effect

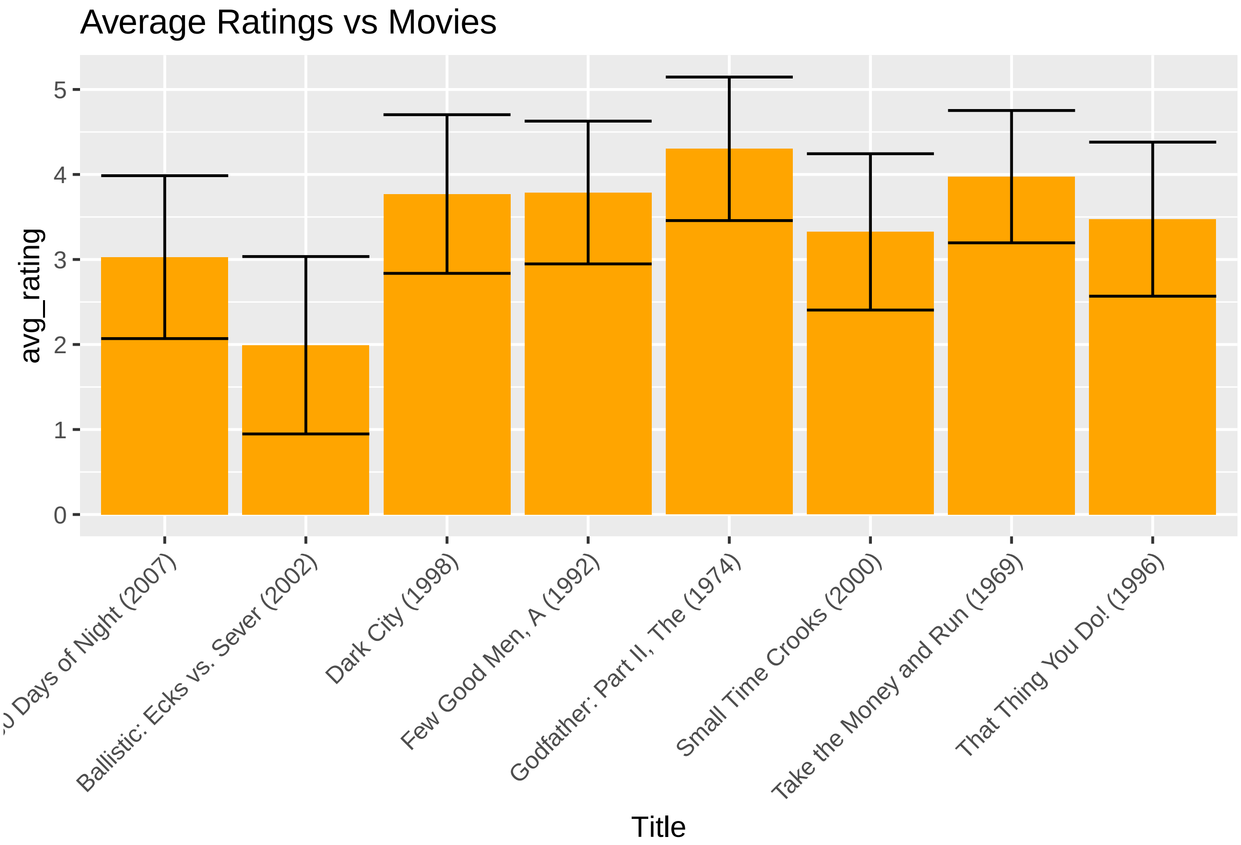
This model is based on data visualization which indicates that some movies are better than others, some users are more generous with their ratings than others and each user has their own preference for different genres.

Data Visualization

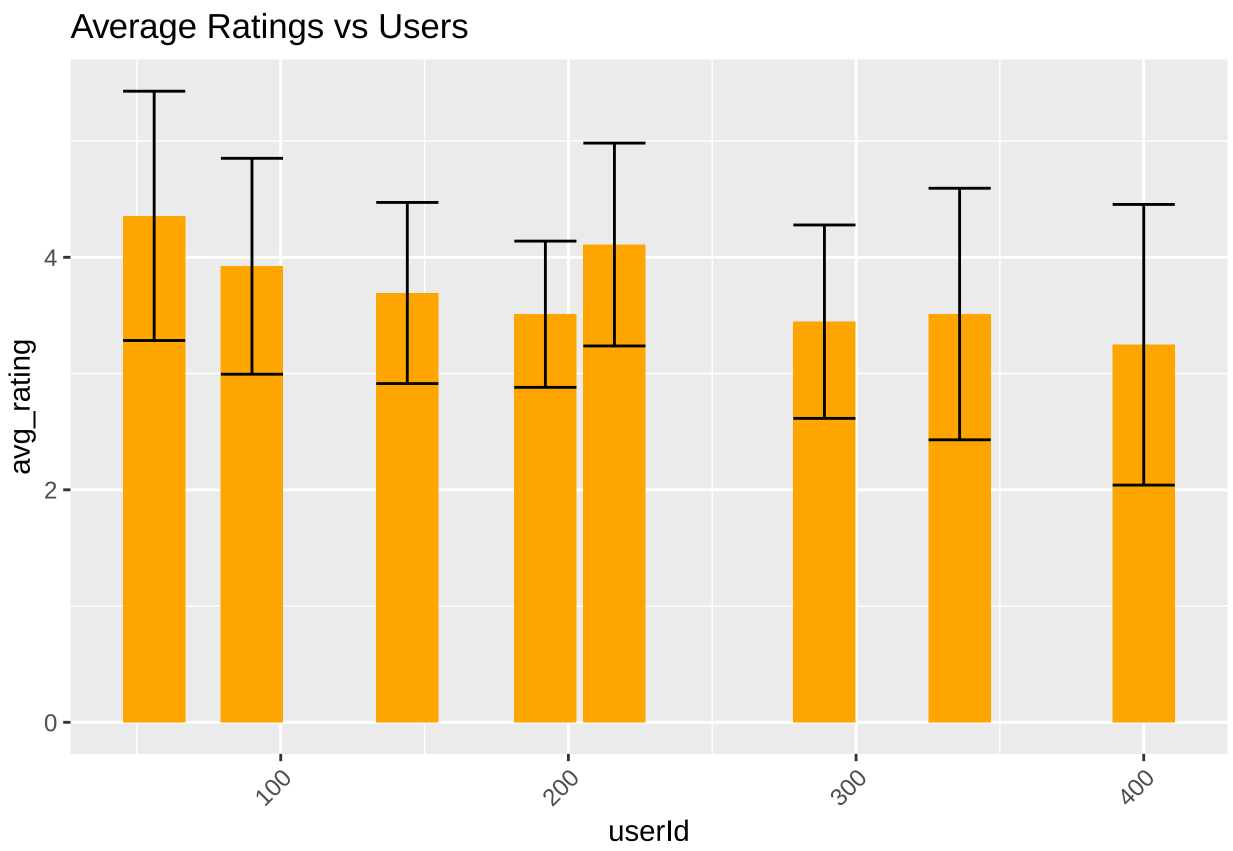
I illustrate the intuition for the model with some examples below.



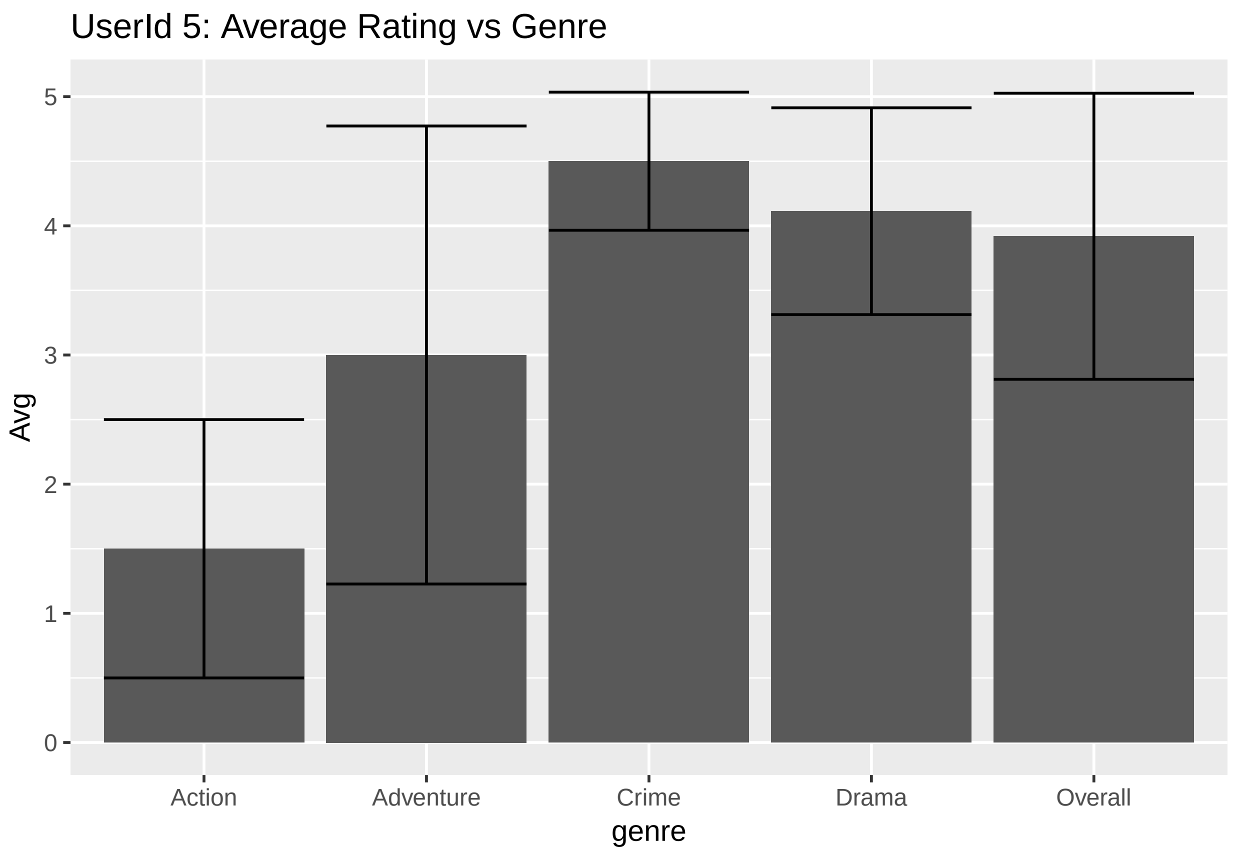
We can see the overall disribution of movies and that movies are generally rated 4 or 3. We see that users are more likely to give round numbers in ratings.



From a random sampling of movies, we can see that some movies are rated higher than others. This makes sense because some movies are better than others.



From a random sampling of users, we can see that some users are more generous than others in rating movies.



We take an example of a user with many entries in the dataset. From this we can see that users may have a preference for certain genres.

This motivates our model.

Data Analysis and Results

The model is based on the belief that the simplest and good guess in each case is simply the mean for that condition.

Without any information about the user or Movie, the best guess is therefore the overall Mean. The **RMSE = 1.061** if we predict the Overall mean

If we account for the Movie Effect, a good approximation is simply the average rating for that Movie by all users. The **RMSE = 0.9439** for this model.

We we account for the user bias, a good approximation is the average of the difference between the rating for a movie by that user and the rating by all other users. The **RMSE = 0.8653** for this model.

To account for the genre preference, a good approximation is to estimate the average of the difference between the rating for a movie in that genre and the ratings by that user for all other movies.

Each movie has multiple genres and the combined effect of all genres is expected to be the average of all genre effects for the genres that are relevant to that movie. The **RMSE = 0.8497** for this model.

Conclusion

Our Model has improved with every layer of complexity. This indicates that our intuition was good. Our RMSE compares favorably with the winners of the Netflix challenge.