MovieLens Project

Abhishek Talari

2023-12-05

Executive summary

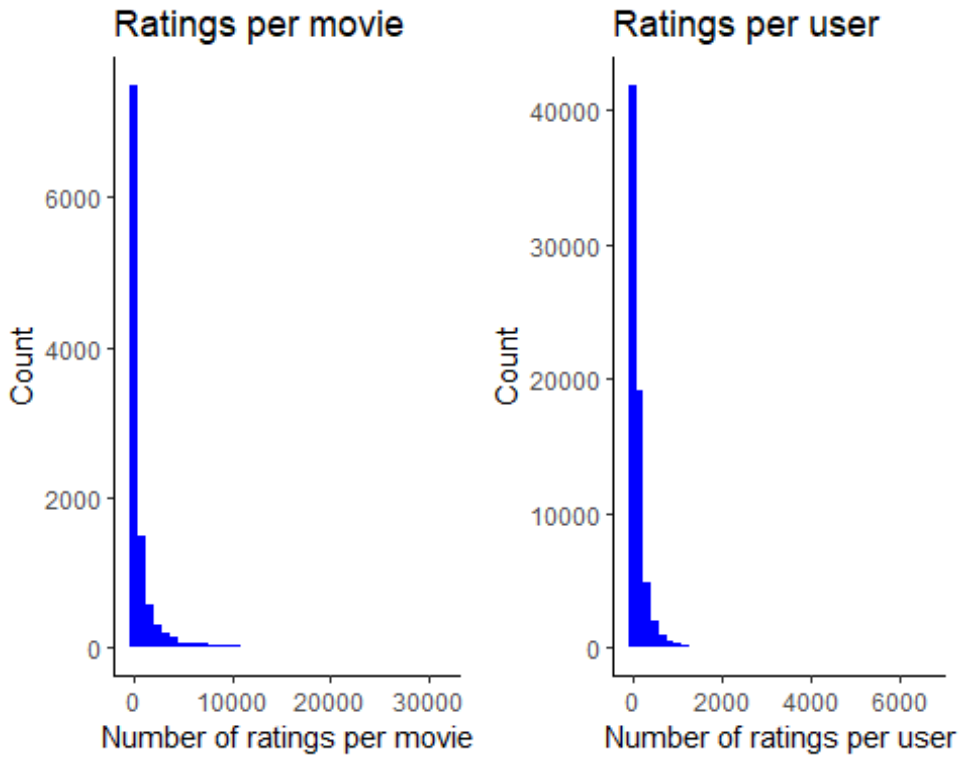
The movielens dataset has over 10 million ratings for more than 10,000 movies given by over 72,000 users. The data set includes the identification of the user, movie, rating, genre, and timestamp.

The goal of this project is to predict movie ratings. To do that, we partitioned data into training set and testing set. We then fit our model with testing set and will calculate the RMSE values. As the dataset is very sparse, we also included regularization in the model.

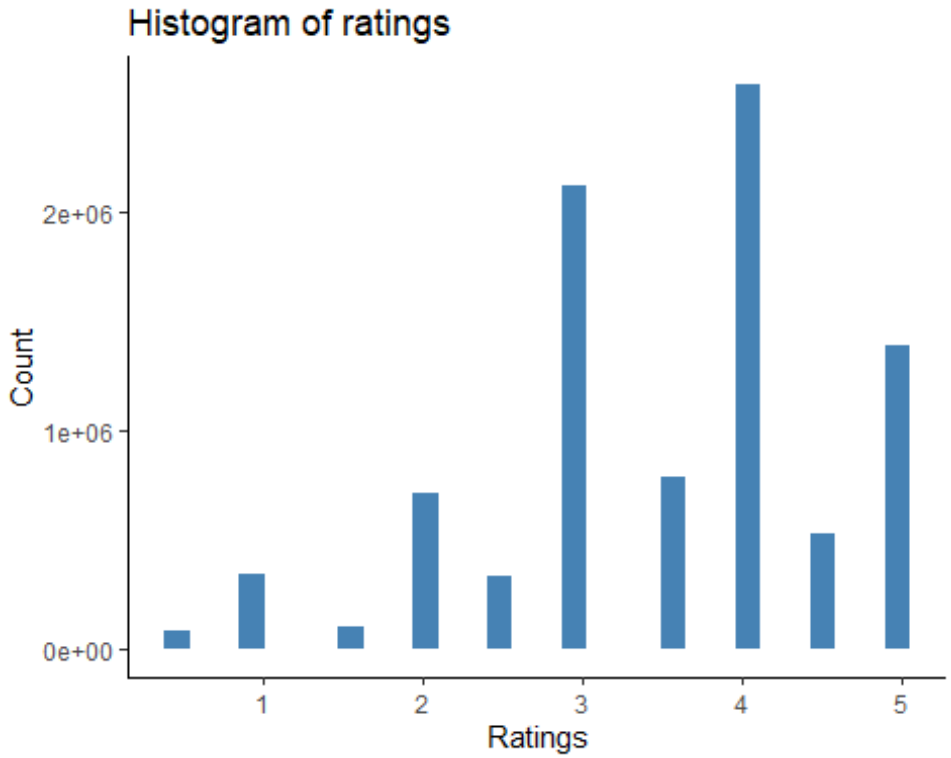
Analysis

Observation 1: From the data set, we see that there are 69878 unique users who have provided ratings and 10677 unique movies were rated. If we have to think about all the possible combinations, there would be more than 746 million combinations for users and movies. But our test set has around 9 million rows that implies that not every user has rated every movie. This number of ratings is only 1.21% of all possible combinations, which necessitates a sparse matrix.

Observation 2: In addition to not having every movie rated by every user, we also see that i) few movies have been rated more than the others and ii) few users have rated more than the others. These can be seen in the two histograms below.



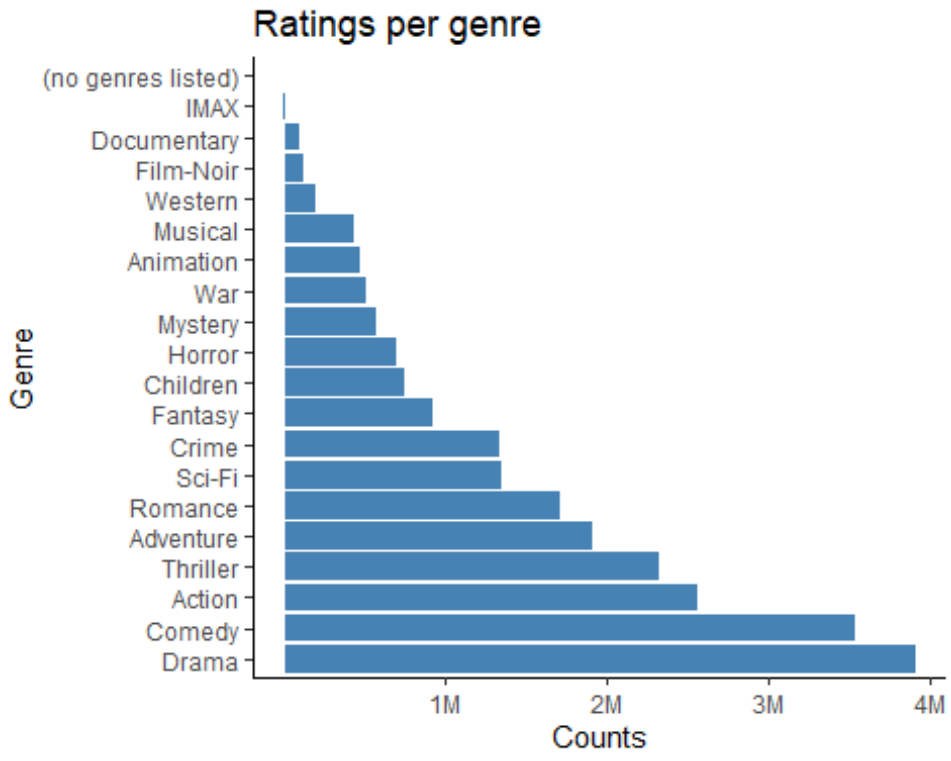
Observation 3: From the below histogram of ratings, users rated more frequently in integers than half-integer and the ratings distribution is left-skewed.

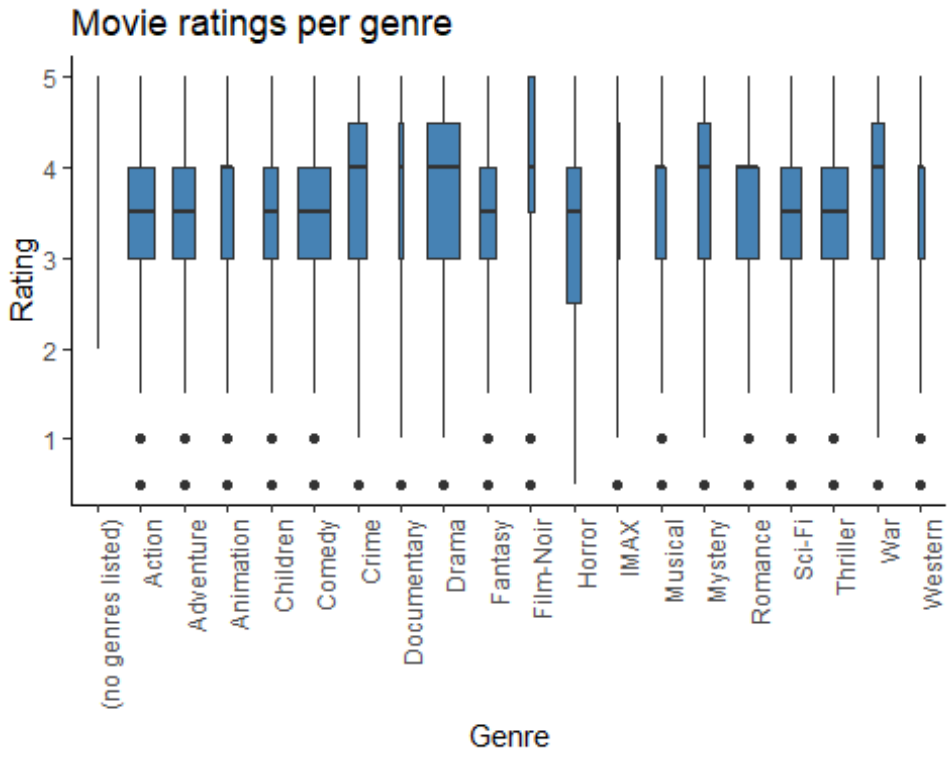


Observation 4: There are 20 different classifications of movie genres. They are as given below:

```
##  [1] "Comedy"             "Romance"            "Action"
##  [4] "Crime"              "Thriller"           "Drama"
##  [7] "Sci-Fi"             "Adventure"          "Children"
## [10] "Fantasy"            "War"                "Animation"
## [13] "Musical"            "Western"            "Mystery"
## [16] "Film-Noir"          "Horror"             "Documentary"
## [19] "IMAX"               "(no genres listed)"
```

Observation 5: It is important to notice that few genres have lot more ratings than others. Drama and Comedy are the most rated genre types. Drama and Film-noir are the better-rated genre types. Horror is the worst rated genre type.





We will be using Root Mean Square Error (RMSE) to measure how close the predictions are to the true values in the validation set.

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

Developing the model:

Here, we first start with the most simple model to have a baseline: predict the same rating regardless of the user, movie or genre.

The model would look like this: $Y_{u,i} = \mu + \epsilon_{u,i}$

Where u is the index for users, i for movies. The estimate for μ is the average of all ratings, which is 3.5124652.

Method	RMSE
Just the average	1.060331

Modelling movie effects:

Factoring the movie effects into the equation to improve the model. We estimate the movie effect as the average of the ratings by a movie.

Method	RMSE
Movie Effect Model	0.9439087

Modelling user effects:

Factoring both movie and user effects to improve the model.We estimate the user effect as the average of the ratings per user.

Method	RMSE
Movie + User Effects Model	0.8651613

Regularization:

Regularization allows us to penalize estimates constructed using sample sizes. The larger the penalty parameter λ, the more the estimate is shrunk. As λ is a tuning parameter, we are doing a grid search to choose its optimal value.

Method	RMSE
Regularized Movie + User Effect Model	0.8651112

Results

To predict movie ratings, we built models that considered movie and user effects. The best model considered all, achieving an RMSE of Regularized Movie + User Effect Model, 0.8651112. The movie effect decreased RMSE the most, suggesting that the movie in itself is of greatest importance to explain the rating.

Method	RMSE
Just the average	1.0603313
Movie Effect Model	0.9439087
Movie + User Effects Model	0.8651613
Regularized Movie + User Effect Model	0.8651112

Conclusion

The project's goal was to predict movie ratings with over 10 million evaluations. To do that, we considered the impact of movies and users. To avoid over-fitting we divided the data set into train and validation. We have computed the RMSE for the models, along with regularization. The best fitted model achieved an RMSE of 0.8651112, which is considered very good as for the course's standards. Hence, we are not proceeding to fine-tune the model further.

The model that we developed did not consider the genre impact. As for the future work, the model can be extrapolated to consider the impact of genres to the ratings. It would have been interesting to have more information about the users (e.g. age and gender) and the movies (e.g. actors, director and language) to try to improve the model.