MovieLens Recommendation System

HarvardX Data Science Professional Certificate

Capstone

Richard Jonyo  
18 February 2022

Contents

[**Introduction** 2](#_Toc95400862)

[**Methods/Analysis** 3](#_Toc95400863)

[**Results** 4](#_Toc95400864)

[**Conclusion** 5](#_Toc95400865)

# **Introduction**

A recommendation system is an information filtering system that attempts to predict the “rating” or “preference” a user would give to an item. The goal of this project was to come up with a recommendation system using the MovieLens datasets which consists of 10 million movie ratings. In this project, we used a smaller subset of the MovieLens which is 10M and contains 10,677 movies by 69,878 users. The dataset has genres, and the movies are rated from 0.5 to 5 with increments of 0.5. A movie can be categoried under a number of genres.

The dataset was split into two sets: a training set (edx) and a final hold-out test set (validation) using code provided by the course. The objective was for the final algorithm to predict ratings with a root mean square error (RMSE) of less than 0.86490 versus the actual ratings included in the validation set.

To develop the recommendation model, a 4-step approach was employed where the initial model asssumed the recommendations abided by a naive mean of all movie ratings. The second model added movie bias to the initial model since some movies are more popular than others. The third model added users bias which improved the RMSE slightly. The final model regularized movie and user biases to improve the RMSE.

Exploratory analysis was conducted on the data using R and R Studio, a language and software environment for statistical computing. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. R Markdown, a component of R Studio was used to compile the report.

# **Methods and Exploratory Analysis**

**Required libraries**

The project utilized and loaded several CRAN libraries to assist with our analysis. The libraries were automatically downloaded and installed during code execution. These included: tidyverse, caret, data.table, and lubridate libraries.

**MovieLens dataset**

The name provided to the subset of the MovieLens dataset is edx dataset which consists of 9,000,055 observations and 6 columns. Each observation is a rating provided by a user for a movie. The dataset contains ratings provided by a total of 69,878 unique users for a total of 10,677 unique movies. Below is a quick summary (top five records) of the dataset:

Table : First 5 rows of edx dataset

![Table

Description automatically generated]()

![Table

Description automatically generated]()

Figure : Training dataset summary

Figure 2 below shows is the distribution of ratings. The average rating in the edx dataset was 3.512465. The minimum rating for a movie was 0.5 and the maximum rating was 5.0

Table : Training set rating distribution

Chart, bar chart, histogram

Description automatically generated

Some movies are more highly rated than others (see Figure 2). Further analysis reveals significant variation in the number of ratings received by each movie (see Figure 3), with the movie with the most ratings, Pulp Fiction (1994), receiving a total of 31362 ratings whereas as many as 126 movies were only rated once. There is clearly a movie effect on the rating awarded and, as such, adjusting for this effect (or bias) was considered worthwhile for inclusion in the training  
algorithm.

According to Figure 2 some movies were rated once while other movies were rated more than 10,000 times

Chart, histogram

Description automatically generated

Figure : Movie ratings by users

Chart, histogram

Description automatically generated

The chart below shows the topmost popular movies in the dataset.

Table : Top 10 Popular Movies

Chart, bar chart

Description automatically generated

# **Results**

# **Conclusion**