Predict movie ratings with MovieLens dataset

Bart Boerman

20-12-2018

Overview of capstone project

[Some text about this capstone project.]

Dataset description

The training dataset provided by HarvardX contains 9000055 records and 6 columns with:

- 69878 users,
- 10677 movies,
- 8316 titles, let's investigate how this can differ from number of movies,
- 707 unieke values in the genres column, muliple genres may apply to one genre,
- 10, ratings, in steps of 0.05 from 0 to 5

The data spreads a timespam from 1995-01-09 11:46:49 until 2009-01-05 05:02:16.

Table 1: Example rows

	userId	movieId	rating	timestamp	title	genres
1	1	122	5	838985046	Bird of Prey (1996)	Action
2	1	185	5	838983525	Bad Moon (1996)	Action|Adventure|Horror
4	1	292	5	838983421	Arsenic and Old Lace (1944)	Comedy|Mystery|Thriller
5	1	316	5	838983392	Some Kind of Wonderful (1987)	Drama|Romance
6	1	329	5	838983392	Field of Dreams (1989)	Drama
7	1	355	5	838984474	Amityville II: The Possession (1982)	Horror

Descriptive statistics

Data wrangling

Exploratory data analysis