

# 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 unique values in the genres column, multiple genres may apply to one genre,
- 10, ratings, in steps of 0.05 from 0 to 5

The data spreads a timespan 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

Make note of the following:

- The title seems to contain the release year of the movie.
- The genre column contains multiple genres separated with a pipe.

## Descriptive statistics

In addition to the quantitative description of the dataset in the previous paragraph some basic statistics about ratings may be of interest.

## Missing values

The dataset has 821944 records with missing values. Which is 9.57 percent. We need to investigate.

Table 2: Percentage missing values per column

userId	movieId	rating	timestamp	title	genres
0	0	0	0	4.57	4.57

What ratings are given to these movies?

Table 3: Ratings on movies with missing meta data

Var1	Freq
0.5	7386
1	7339
1.5	8712
2	20077
2.5	30350
3	60875
3.5	81818
4	105292
4.5	56248
5	32875

The average rating is 3.5228093, which is almost equal to the overall average rating: 3.5124652.

Let's not ignore these cases.

## Data wrangling

Feature engineering

- year, month, day of week, hour
- extract release year from title
- one-hot encode genre
- add number of ratings per movie
- add number of rating per user

## Exploratory data analysis