Movielens RMD Report

Knut Ulsrud

6/3/2020

Overview

This section presents the project's goals and a high-level overview of project data.

Goal of the project and key steps that were performed

This project's goal was to train a movie recommendation system based on Movielens data. This report was produced as a submission to the Data Science: Capstone course provided by HarvardX on the EdX platform.

Dataset Description

The Movielens dataset that was provided through this course has 9,000,055 rows and six columns. Columns contain the following information:

Variable name	Variable Description
1. userId	A unique numeric value assigned to each user u that has provided a rating
2. movieId	A unique value assigned to each movie m, in numeric format
3. rating	The rating provided to a specific movie m by a user u, in numeric format
4. timestamp	the time at which the rating was provided, in numeric format
5. title	The name of each movie m, in character format
6. genres	The genre combinations of each movie, in character format

The following table shows the first four rows of the dataset.

	userId	movieId	rating	timestamp	title	genres
1	1	122	5	838985046	Boomerang (1992)	Comedy|Romance
2	1	185	5	838983525	Net, The (1995)	Action|Crime|Thriller
4	1	292	5	838983421	Outbreak (1995)	Action|Drama|Sci-Fi|Thriller
5	1	316	5	838983392	Stargate (1994)	Action|Adventure|Sci-Fi

Methods and Analysis

This section presents steps taken to perform initial analysis on the dataset. It includes the following sections: Data Cleaning, Data Exploration and Visualization, Insights Gained, and Modelling Approach.

Data Cleaning

A high-level check revealed that there were no missing values in the Movielens dataset provided for this assignment. The data was already presented in a tidy format (one row equals one observation) which means no additional transformation was required to start using the dataset.

Data Exploration and Visualization

It is likely that there is systematic variation between users and their rating, specific movies and their rating, and a movies genre and the movie's rating. This section explores the relationships between these features to understand ones are likely to successfully predict movie ratings.

Movie Effects

This section explores the possibility of movie effects, and considers the relationship between a movie's rating and the number of ratings it has received.

title	rating	n
Blue Light, The (Das Blaue Licht) (1932)	5.00	1
Fighting Elegy (Kenka erejii) (1966)	5.00	1
Hellhounds on My Trail (1999)	5.00	1
Satan's Tango (SÃ¡tÃ¡ntangÃ³) (1994)	5.00	2
Shadows of Forgotten Ancestors (1964)	5.00	1
Sun Alley (Sonnenallee) (1999)	5.00	1
Constantine's Sword (2007)	4.75	2
Human Condition II, The (Ningen no joken II) (1959)	4.75	4
Human Condition III, The (Ningen no joken III) (1961)	4.75	4
Who's Singin' Over There? (a.k.a. Who Sings Over There) (Ko to tamo peva) (1980)	4.75	4

Displaying movies with the highest average rating we see that all of the top movies have few ratings assigned to them.

title	rating	n
Accused (Anklaget) (2005)	0.5000000	1
Besotted (2001)	0.5000000	2

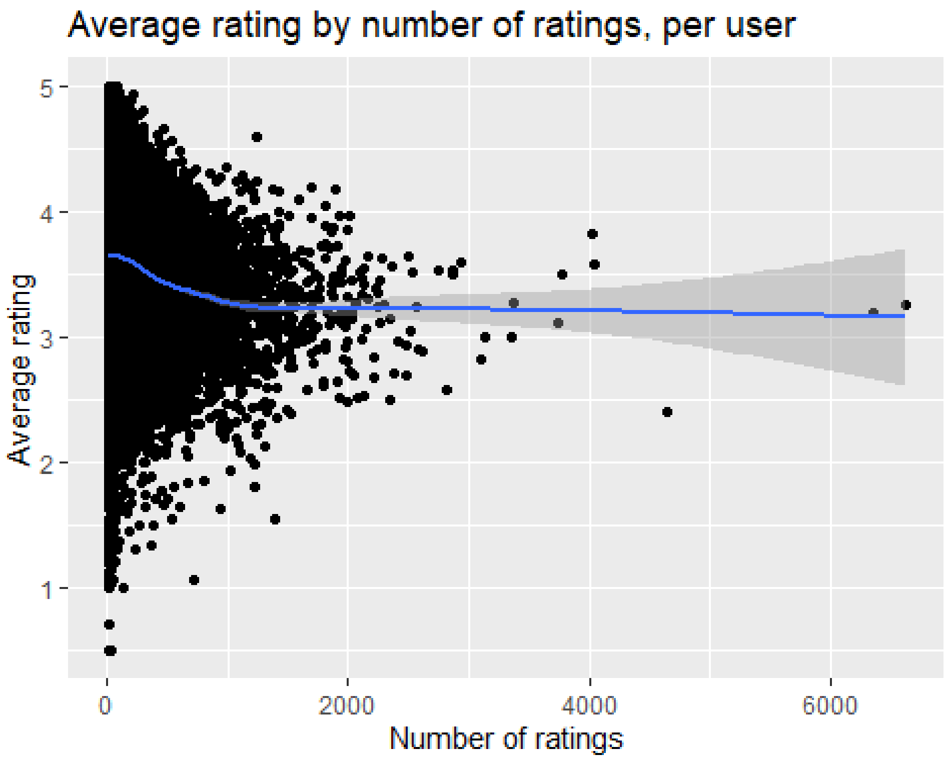
Confessions of a Superhero (2007)	0.5000000	1
Hi-Line, The (1999)	0.5000000	1
War of the Worlds 2: The Next Wave (2008)	0.5000000	2
SuperBabies: Baby Geniuses 2 (2004)	0.7946429	56
Hip Hop Witch, Da (2000)	0.8214286	14
Disaster Movie (2008)	0.8593750	32
From Justin to Kelly (2003)	0.9020101	199
Criminals (1996)	1.0000000	2

Displaying movies with the lowest average rating we see that most of the bottom movies have few ratings as well.

The insight we get from displaying the best and worst movies is that we have confirmed the intuition that ratings vary by movie, and that movies with very high or very low scores have few ratings. The latter finding indicates that regularizing can improve the model's predictions.

User Effects

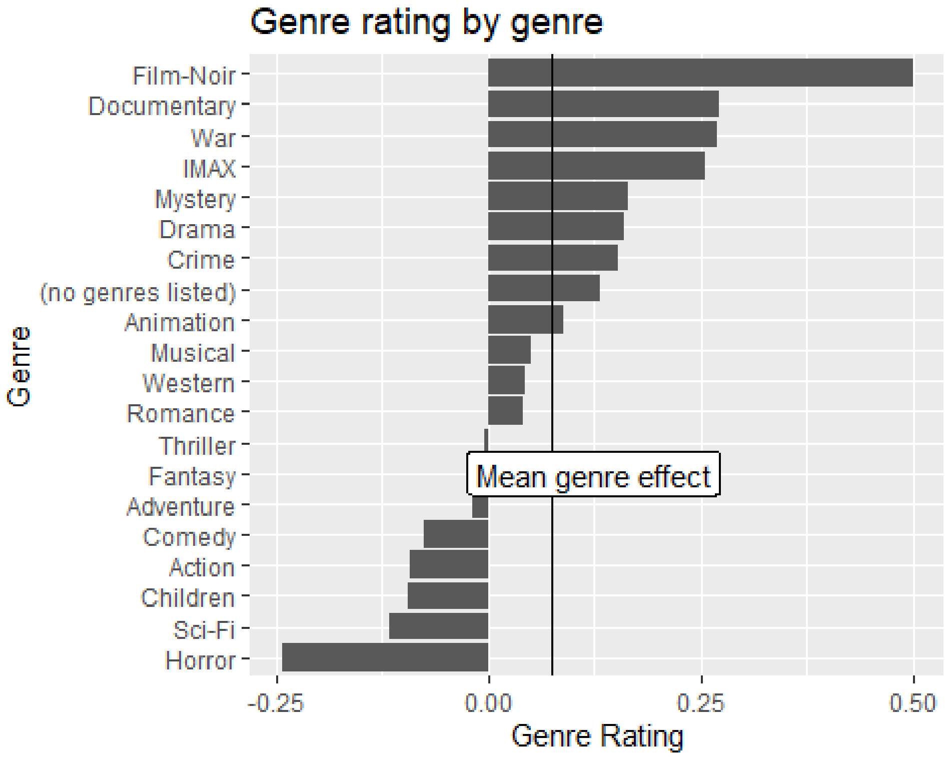
This section explores the possibility of user effects, and considers the relationship between users' average movie ratings and the number of ratings they have provided.



The graph shows that some users have submitted more ratings than others. It also shows that users with fewer ratings are likely to have more extreme (higher or lower, with bias towards higher) average ratings. The latter finding indicates that regularizing can improve the model's predictions.

Genre Effects

This section explores the possibility of genre effects.



The graph shows that different genres systematically deviate from the mean overall mean. While movies are assigned combinations of genres, this graph shows the effect from isolated genres. Aggregate genre effects can be assigned to movie ratings based on the movie's genre combination.

Insights gained

There is systematic variation between movies scores, between genre scores, and between user ratings. These are three variables that we can use to predict movie scores in the validation set. The analysis also shows that movies with few ratings, and users with few ratings, sometimes tends towards more extreme values. This means that regularization will likely improve our modelling results.

Modeling approach

The model will follow the following formula to predict movie scores.

$Y_{u,i} = \mu + b_i + b_u + \sum k = 1^K x_{u,i} \beta_k + \varepsilon_{u,i}$, with $x_{u,i}^k = 1$ if $g_{u,i}$ is genre k

It will calculate movie effects b_i, user effects b_u, and genre effects k that summarize genre scores for all genres that apply to a particular movie rating by a user.

Both movie and user effects are regularized to reduce the contribution from observations with low n. The regularized variables are defined as follow:

$b_i = sum(rating - \mu - \sum k)/(n + \lambda)$, and $b_u = sum(rating - \mu - b_i - \sum k)/(n + \lambda)$. This means that b_i captures residuals per movie after first estimating the average rating and the genre effect. b_u captures residuals per user after first estimating the average rating, the genre effect, and movie effects.

λ or lambda is the regularization parameter. It was tuned by testing values from 0 to 10 at progressively smaller increments (0.1 at the smallest) optimizing between the estimated movie effect and user effect.

Results

A movie recommendation system based on the presented approach produces the following RMSE:

Regularized movie and user effects + genre effects: 0.8647843

Additional exploration could be done on the effects from timing of rating and movie release date. This could potentially yield additional improvements to the model's RMSE.