Capstone 2 Milestone Report: Goodreads Recommendations

Eitan Angel

July 15, 2019

Contents

List of Figures

List of Tables

1 Introduction

1.1 Problem: Make Book Recommendations for Goodreads Users

Goodreads is a social site for readers and for book recommendations. In this project we make recommendations to existing users of books they would most enjoy which they have not yet rated. To do so, we use a collaborative filtering filtering approach and compare the error in our recommendations to the error of some baseline models. Once we have made this model it is not so difficult to provide recommendations to new users who are willing to rate a few books.

1.2 Data: Goodbooks-10k

This is a dataset scraped from Goodreads of the 10,000 most popular books (by number of ratings). It contains book ratings by over 50,000 users, as well as user-created tags, including books tagged "to-read" and considerable data on the books themselves in both a `.csv` file and in an archive of `.xml` files. The basic model will only consider the explicit book ratings although a next step is to find implicit relationships, say among tags and users or books.

1.3 Approach: Collaborative Filtering via Matrix Factorization

We will use a Funk SVD-like collaborative-filtering approach. First we create a user-book matrix of ratings V (sparsity $\approx 99\%$). Following that, we can use Non-negative Matrix Factorization (NMF) to find matrices W and H which decompose V as $V \approx WH$ by minimizing a root-mean-square error (RMSE) between V and WH.

Consider W to be matrix of latent user features and H to be a matrix of latent book features. By matrix completion, we mean to consider the matrix $A = WH$ as "filling in" those ratings which are blank in V. To make recommendations for a user, return the top-N values in the row of A corresponding to that user (which they have not already rated). We can compare the RMSE matrix factorization techniques to various simpler baseline models.

2 Exploratory Data Analysis

While the dataset has considerable features and metadata on books and tags, we will focus on ratings. The three relevant files are `books.csv`, `ratings.csv`, and `to_read.csv`.

2.1 Books

The file `books.csv` has a row for each of the 10,000 most rated books on Goodreads and the following 23 columns: `book_id, goodreads_book_id, best_book_id, work_id, books_count, isbn, isbn13, authors, original_publication_year, original_title, title, language_code, average_rating, ratings_count, work_ratings_count, work_text_reviews_count, ratings_1, ratings_2, ratings_3, ratings_4, ratings_5, image_url, small_image_url` .

We will inspect whether `average_rating` is influenced by other `books.csv` features, as well as some of the top-rated books, oldest books, most- and least-reviewed books

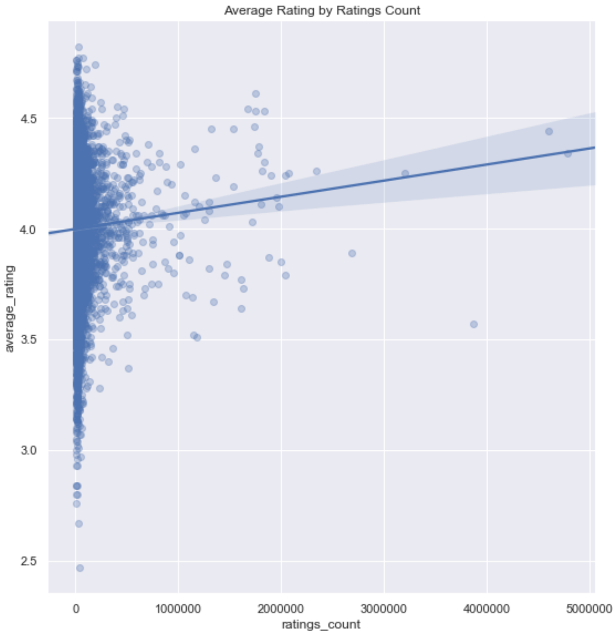

Figure 1: There is some effect of `ratings_count` on `average_rating` – more popular books are better rated.

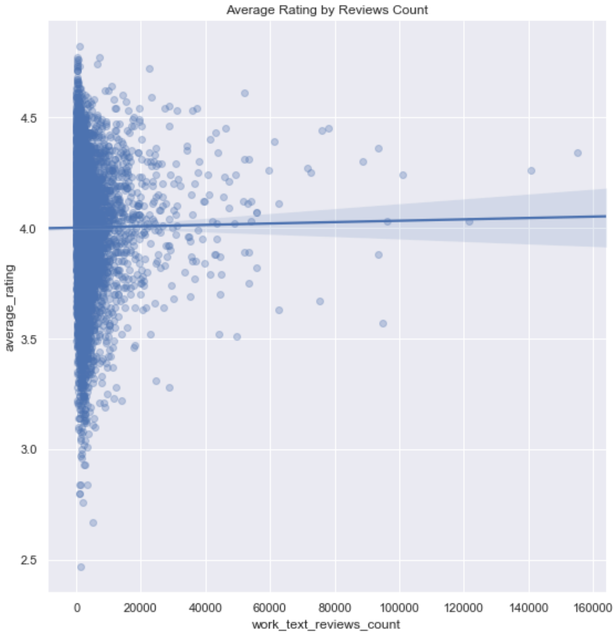

Figure 2: The number of reviews does not have a significant effect on `average_rating`.

authors	title	avg_rating	ratings
Suzanne Collins	The Hunger Games (The . . .	4.34	4942365
J.K. Rowling, Mary GrandPré	Harry Potter and the Sorcerer’s. . .	4.44	4800065
Stephenie Meyer	Twilight (Twilight, #1)	3.57	3916824
Harper Lee	To Kill a Mockingbird	4.25	3340896
F. Scott Fitzgerald	The Great Gatsby	3.89	2773745
John Green	The Fault in Our Stars	4.26	2478609
Veronica Roth	Divergent (Divergent, #1)	4.24	2216814
J.R.R. Tolkien	The Hobbit	4.25	2196809
Jane Austen	Pride and Prejudice	4.24	2191465
J.D. Salinger	The Catcher in the Rye	3.79	2120637

Table 1: The most popular books on Goodreads.

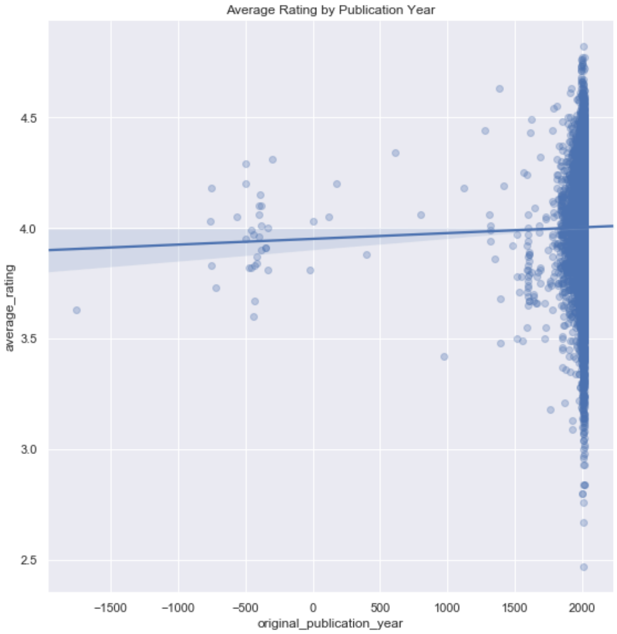

Figure 3: The effect of `original_publication_year` on `average_rating` is not significant. Negative values are books published 1 BCE or earlier.

authors	title	average_rating
Bill Watterson	The Complete Calvin and Hobbes	4.82
J.K. Rowling, Mary GrandPré	Harry Potter Boxed Set, Books 1-5	4.77
Brandon Sanderson	Words of Radiance (The Stormlight ...	4.77
Francine Rivers	Mark of the Lion Trilogy	4.76
Anonymous ...	ESV Study Bible	4.76
Bill Watterson	It's a Magical World: A Calvin and ...	4.75
Bill Watterson	There's Treasure Everywhere: A Calvin ...	4.74
J.K. Rowling	Harry Potter Boxset (Harry Potter, #1-7)	4.74
J.K. Rowling	Harry Potter Collection (Harry Potter, #1-6)	4.73
Bill Watterson	The Indispensable Calvin and Hobbes	4.73

Table 2: Calvin & Hobbes and Harry Potter dominate the average ratings.

authors	year	title
Anonymous...	-1750.0	The Epic of Gilgamesh
Homer, Robert Fagles ...	-762.0	The Iliad/The Odyssey
Homer, Robert Fagles	-750.0	The Iliad
Anonymous ...	-750.0	The I Ching or Book of Changes
Homer, Robert Fagles ...	-720.0	The Odyssey
Aesop, Laura Harris ...	-560.0	Aesop's Fables
Anonymous, Juan Mascaró	-500.0	The Upanishads: Translations from the Sanskrit
Sun Tzu, Thomas Cleary	-500.0	The Art of War
Anonymous ...	-500.0	The Dhammapada
Confucius, D.C. Lau	-476.0	The Analects

Table 3: The oldest books in the dataset.

authors	title	avg	count	ratio
Cynthia Hand, Brodi Ashton, ...	My Lady Jane (The Lady ...	4.12	12794	0.274
Amie Kaufman, Jay Kristoff, ...	Gemina (The Illuminae ...	4.56	10960	0.265
Amie Kaufman, Jay Kristoff	Illuminae (The Illuminae ...	4.32	44500	0.264
Angie Thomas	The Hate U Give	4.62	32610	0.236
Stephanie Garber	Caraval	3.97	30975	0.233
Marissa Meyer	Heartless	4.06	33348	0.233
Sarah Pinborough	Behind Her Eyes	3.77	17944	0.231
Julianne Donaldson	Edenbrooke (Edenbrooke ...	4.34	28536	0.229
Pam Muñoz Ryan	Echo	4.36	14864	0.225
Victoria Schwab	This Savage Song (Monsters ...	4.14	17210	0.225

Table 4: The ratings ratio is `work_text_reviews_count` divided by `work_ratings_count`. The majority of the greatest ratings ratio books are romance novels.

authors	title	avg	count	ratio
Cynthia J. McGean	Henry & Ramona	4.14	11106	0.000270
John D. Rateliff, J.R.R. Tolkien	The History of the Hobbit, Part One…	3.81	108399	0.000424
Frank Miller	Sin City: Una Dura Despedida …	4.21	9115	0.000439
Janet Evanovich	Janet Evanovich Three and Four ….	4.34	63691	0.000612
Dean Koontz, Leigh Nichols	Cold Fire / Hideaway / The Key to …	4.16	17581	0.000626
Mark Cotta Vaz	The Twilight Saga Breaking Dawn …	4.30	188136	0.000712
Richard Lancelyn Green, …	The Further Adventures of Sherlock …	4.40	36863	0.000976
Amazon	Kindle Paperwhite User's Guide	3.72	15002	0.001037
John Williams	Harry Potter and the Chamber of …	4.61	29409	0.001054
Jenö Barcsay	Anatomy for the Artist	3.97	21640	0.001107

Table 5: Books with the least ratings ratio.

Table 6: `ratings.csv` and `to_read.csv`

user_id	book_id	rating
1	258	5
2	4081	4
2	260	5
2	9296	5
2	2318	3

(a) `ratings.csv` consists of 5,976,479 entries, 53,424 users, and 10,000 books.

user_id	book_id
9	8
15	398
15	275
37	7173
34	380

(b) `to_read.csv` consists of 912,705 entries, 48,871 unique `user_ids`, and 9,986 unique `book_ids`.

2.2 Ratings

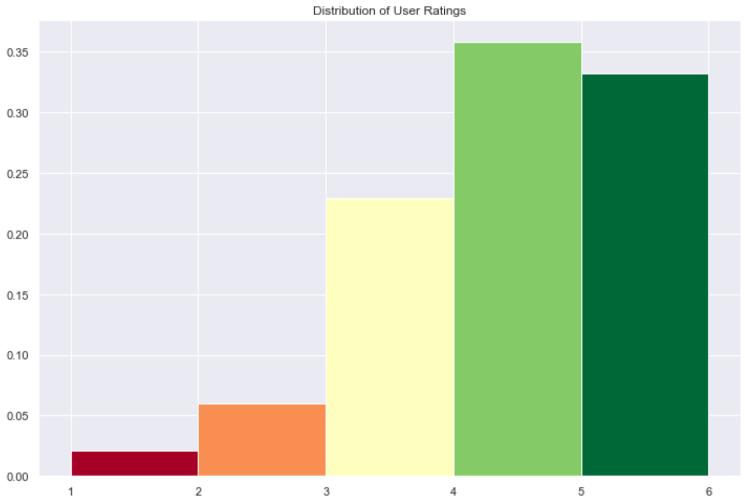

Figure 4: Ratings of 4 or 5 are by far most common.

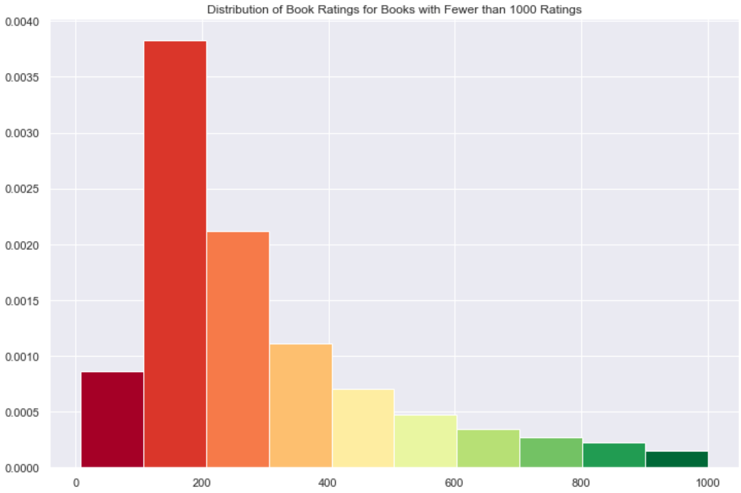

Figure 5: The distribution of ratings by book in `ratings.csv` is left skew. The range is 8–22806 though the interquartile range is 155–503. Since the tail is long we plot the distribution for books with fewer than 1000 ratings.

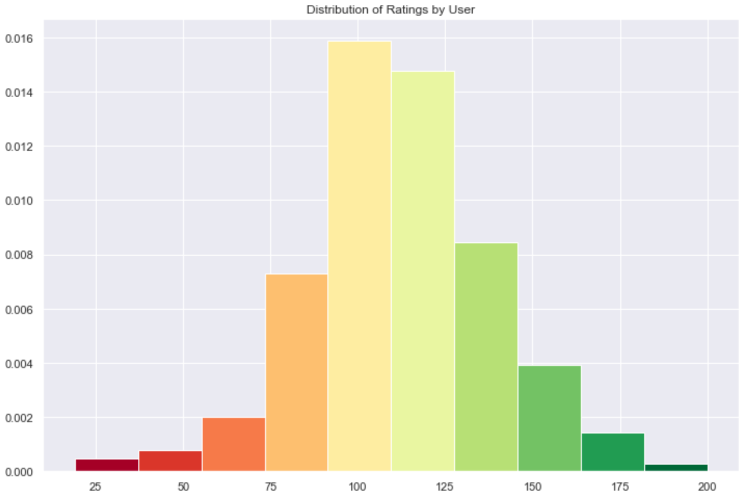

Figure 6: The range of reviews by user is 19–200.

2.3 To-Read

While there are other tags, to-read is the only tag we will consider in this version of the model. We can optionally let users decide against recommendations of books in this list. Most users tag at least one book to-read and almost all books are tagged to-read by some user.