Personalized Book Recommendation System

Rachit Jain, Bibhabasu Das, Shreya Gupta, Abhranil Chakrabarti

December 2022

1 Overview

In this digital era, most industries have seen a major shift towards using digital platforms to push their offerings, with a sharp rise in user traction during the COVID-19 pandemic. User experience is key on such platforms and a smart equivalent of a traditional salesperson becomes increasingly essential. This is where recommendation systems come into the picture. They understand user behavior and their requirements to recommend them the offerings/products that are most relevant to them, and hence are most likely to buy. Some of the major e-commerce players in the industry, like Amazon, have figured out how to create the best recommendation system, that keeps improving with time, owing to increased information availability for each user. The purpose of this project is to explore recommendation methods on a specific application, that can be easily extended to any domain, displaying the power and relative ease of implementation of such methods.

2 Objective and Scope

This project aims to build a recommendation system that uses historical book rating information available for users to recommend books to them. The key idea behind a recommendation system is that users who share similar preferences are most probable to like similar items; users who like an item are probable to like a similar item.

3 Data Used

We rely on data retrieved from multiple data sources to get information on books, ISBN, author, review, and corresponding reviewer information. The following data sources are used:

- Amazon Data: 51.3M reviews and 2.9M products
  - **Sample Review:** ReviewerID, ProductID, Review, Overall Rating, Review Date, etc.
  - **Meta Data:** ProductID, Title, Price, Also Bought, Also Viewed, Bought Together, Category, etc.
- **Kaggle Data:** 271k books, 279k Users, 1.1M Reviews
  - **Book Data Labels:**ISBN, Book-Title, Book-Author, Publisher, Year-of-Publication, Image-URL
  - **Rating Data Labels:** UserID, ISBN, Rating; User Data: User-ID, Location, Age

4 Data Cleaning & Preprocessing

We use Kaggle data as our base and extract information from Amazon dataset to enhance the models to be built. We perform the following data cleaning steps:

- **Filtering data:** Since the user-book rating information is very sparse (i.e., ratings available for a very small subset of books for each user), we limit our recommendation system to rely on information for users who have rated at least 200 books and for books that have at least 50 ratings. This makes the ratings as well as the users more "reliable".
- **Extracting book meta-data:** The Kaggle dataset does not contain enough information about books to get a reliable understanding of user preferences. So, we extract additional data on books like 'Description', 'Price' and 'Category'. We also get data on behavior of users who bought those books, like 'Also Bought' and 'Also Viewed'.

5 Understanding the Data

The following charts indicate how the data is distributed. In Figure 1, we can see that the 3 most frequently rated authors are Agatha Christie, William Shakespeare and Stephen King. This is not surprising as these are some of the most widely known authors.
The distribution of ratings indicates that people generally rate a book only if they moderately liked it or really liked it. Very few users bother to rate books if they completely disliked it.

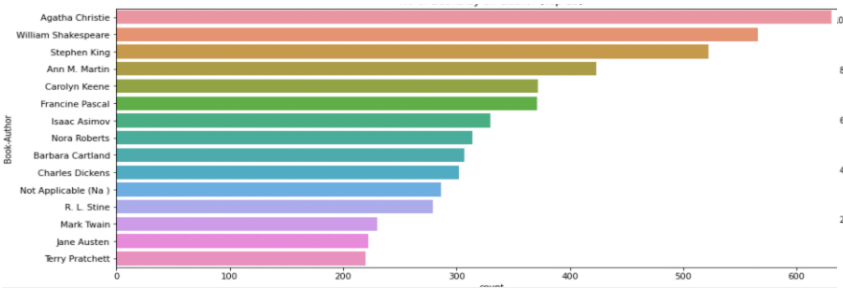

Figure 1: Most Frequently Rated Authors

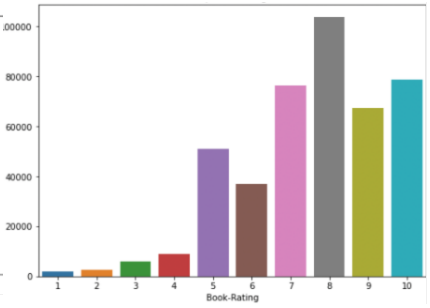

Figure 2: Distribution of Ratings

6 Recommendation Approach

The system makes different recommendations based on the kind of user.

- **New User**: For a new user, we recommend the most popular book in general or within a specified genre. This is because we have no data on the user to determine the user's preferences. In a real-time system, the system would recognize this as a new user, and start tracking information for them to be able to make better and more relevant recommendations in the future.
- **Existing User:** Since we already have the ratings for some books in our data, we rely on this information to provide book recommendations (both at an overall and at a genre level)

Similarity Measure As the name suggests, similarity measures indicate how "similar" or "close" two users or two items are to each other. One of the most common similarity measures is **"cosine similarity"**, generated by mapping features available for the user/item in a latent space.

We use some of the most widely applied methodologies for recommendation systems and compare the results obtained from each.

- **Naïve Recommendation:** Recommend the most popular book for new users, i.e., rank books by their ratings and then recommend
- **Content Based Recommendations System:** Build models that rely on user and book attributes as inputs to predict ratings. Key techniques used: Linear Regression, Holistic Regression, XGBoost

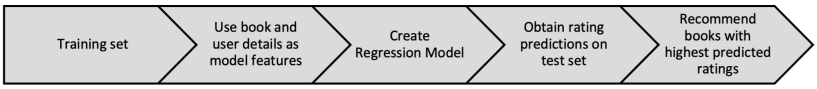

Figure 3: Content-based Filtering Method

- **Collaborative Filtering:** The dataset used for recommendation systems is generally very sparse (has a lot of blanks). This is because we have information available for very few items for each user. Collaborative filtering deals with this sparsity by decomposing the data lower dimensional matrices that represent the users and items (in a latent space). The key idea is that people who agreed in their evaluation of certain items are likely to agree again in the future.
- **Hybrid Methods:** Collective matrix factorization to combine user-item ratings with book and user attributes to give more informed recommendations

7 Optimization for Recommendation

For recommendations to a particular user, we solve the following optimization problem to account for the liking of the user along with the variety of the recommended book set
Variables:

- R_i - Predicted rating of book i by the chosen user
- B_i - Feature vector of book i
- X_i - Decision Variable: 1 if book i is recommended and 0 otherwise

- Y_{ij} - Decision Variable: 1 if book i and book j are both in the recommended list and 0 otherwise

Objective:

$$max_X \sum_i X_i * R_i$$

Constraints:

- We must recommend K books (K=5 in our case):
  $\sum_i X_i = K$
- Recommended books must not be very similar ($\rho = 0.6$ in our case):
  $Y_{ij} * \frac{B_i^T B_j}{|B_i||B_j|} <= \rho, \forall i, j$
- Linking Constraints:
  $Y_{ij} >= X_i, Y_{ij} >= X_j \forall i, j$
- Binary Decision Variables:
  $X_i \in \{0, 1\}; Y_{ij} \in \{0, 1\}$

We perform the optimization over the 100 most highly rated books of the user to make the problem more tractable

8 Results

We assess the performance of the models on the root mean squared error on the test set predictions. This is compared against a baseline whose predicted rating for a user-item pair is the average rating with added Gaussian noise. XGBoost and Collaborative Filtering (via SVD) perform the best on the test set, achieving a reduction of 35% and 33% over the baseline

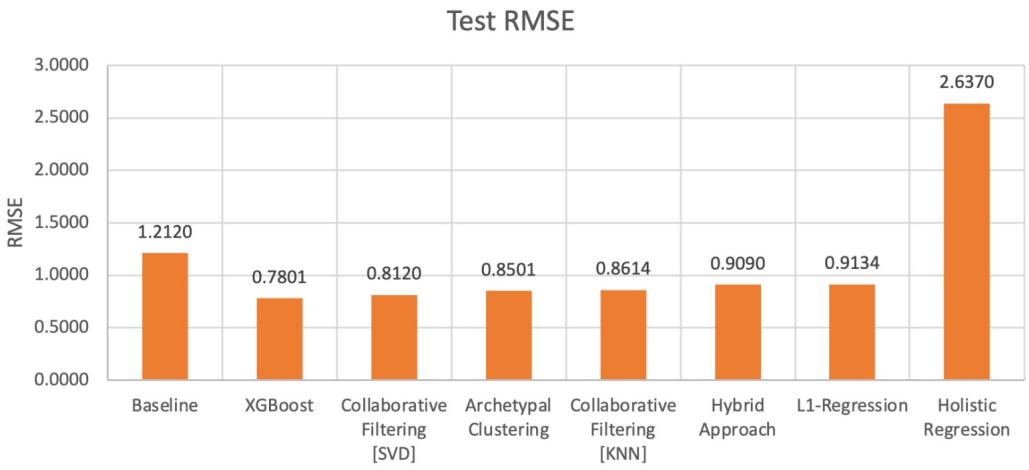

Figure 4: Test Set Results

We also look at 2 users, looking at their ratings to understand their preferences and assess the recommendations provided to each of them

8.1 User 9512 (most number of ratings)

Archetypal Clustering		XGBoost		SVD with Optimization		Hybrid Filtering	
Category	Title	Category	Title	Category	Title	Category	Title
Science Fiction & Fantasy	A Dance with Dragons	Mystery	The Fury A Henry Parker Novel	Literature & Fiction	The Winds of War	Travel	A Thousand Days in Venice Ballantine Readers Circle
Literature & Fiction	This Perfect Day	Mystery	Edge of Danger	Biographies & Memoirs	Quartered Safe Out Here	Travel	The Vintage Caper
Politics & Social Sciences	Three Cups of Tea	Mystery	The Bordeaux Betrayal A Wine Country Mystery Wine Country Mysteries	Literature & Fiction	The Long Ships	Travel	A Nail Through the Heart A Novel of Bangkok
Romance	Music of the Heart	Mystery	A Stranger Like You A Novel	Religion & Spirituality	Survival in Auschwitz	Parenting & Relationships	A Boy Should Know How to Tie a Tie And Other Lessons for Succeeding in Life
Crafts	On Paper The Everything of Its TwoThousandYear History	Mystery	Cold Case Alan Gregory	Cookbooks	Dr. Atkins New Diet Revolution	Travel	City of Dark Magic A Novel City of Dark Magic Series

Figure 7: Recommendations for User 9512 across different methods

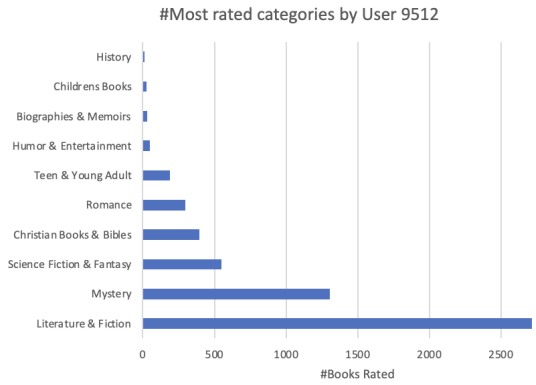

Figure 5: Most rated categories

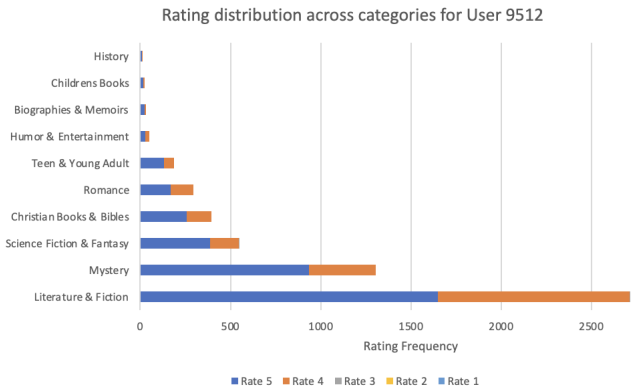

Figure 6: Distribution of Ratings

- User 9512 has mostly rated books in 'Literature & Fiction', 'Mystery' and 'Science Fiction & Fantasy' categories. The rating distribution indicates that the user rates only those books that they like (as all ratings are either 4 or 5).
- We see that XGBoost and Archetypal Clustering give the best results for this user, as the categories recommended by them are closest to the most rated categories by this user.
- The optimization approach does recommend a diverse set of books while the Hybrid approach recommends books in categories which the user has not rated

8.2 User 53 (randomly selected)

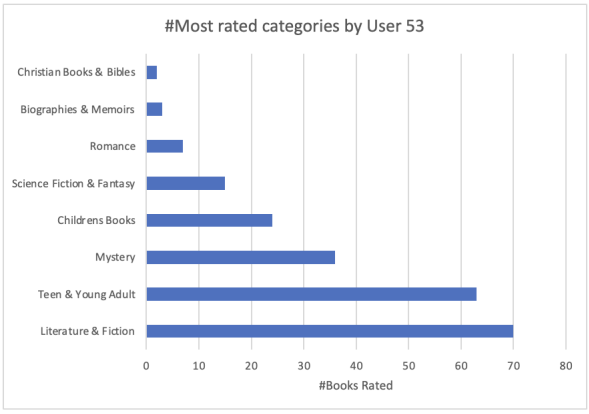

Figure 8: Most rated categories

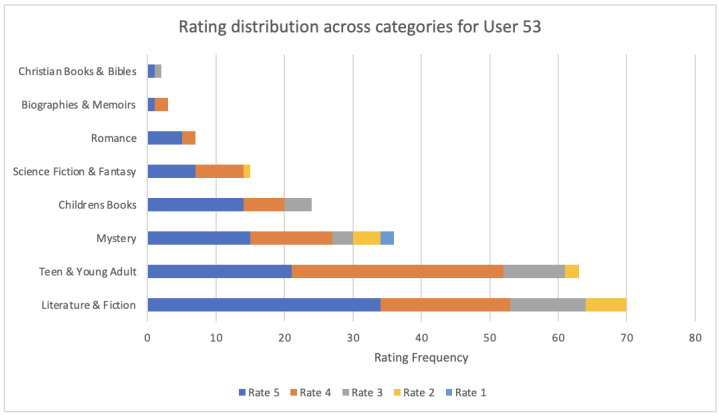

Figure 9: Distribution of Ratings

Archetypal Clustering		XGBoost		SVD with Optimization		Hybrid Filtering	
Category	**Title**	**Category**	**Title**	**Category**	**Title**	**Category**	**Title**
Literature & Fiction	Folly	Childrens Books	The Journal of Curious Letters The 13th Reality	Romance	Seduced By Shadows A Novel of the Marked Souls	Childrens Books	Life as We Knew It
Literature & Fiction	Florence of Arabia A Novel	Teen & Young Adult	Incarceron	Romance	Blood on Silk An Awakened By Blood Novel	Childrens Books	The Line
Literature & Fiction	Trespass A Novel	Teen & Young Adult	The Summoning	Romance	Immortal Warrior	Childrens Books	Dead Is a Battlefield
Childrens Books	Click Clack Moo Cows That Type	Teen & Young Adult	The Awakening Darkest Powers	Cookbooks	The GoodtoGo Cookbook	Childrens Books	Trash
Science Fiction & Fantasy	Spin State	Literature & Fiction	The Lucky One	History	The Murder of King Tut The Plot to Kill the Child King	Childrens Books	Dark Secrets 1 Legacy of Lies and Dont Tell

Figure 10: Recommendations for User 53 across different methods

- User 53 has very few ratings (225 ratings) in comparison to user 9512 (5617 ratings)
- User 53 has mostly rated books in 'Literature & Fiction', 'Teen & Young Adult' and 'Mystery' categories. The rating distribution indicates that the user rates books they liked as well as the ones they didn't like.
- We see that XGBoost and Archetypal Clustering give the best results for this user, as the categories recommended by them are closest to the most rated categories by this user.

- SVD with optimization provides recommendations in newer categories for which the user has rated very few books (or none at all). Hybrid Filtering recommends all books in the same category ('Children's books), which despite having a few books rated by the user, might not be the best recommendation (no diversification and not a good representation of user rating behavior).

9 Impact

This project is an implementation of some of the most widely used recommendation practices across industries, and can be extend to the following applications:

- The recommendation system can be used by a marketplace as a part of its push strategy to understand its users, recommend them books while also deciding price point of these books dynamically (based on trending prices in the book category and user willingness to pay historically).
- This system is agnostic to product category and can be embedded as a part of customer experience improvement for any platform, thereby empowering the customer with reliable and personalized information