**Connecting a Chatbot to an AI-as-a-Service**

Alan L. Dennis

University of the Cumberlands

**Book Recommendation System**

A recommendation engine is a class of machine learning which offers relevant suggestions to the customer. A recommendation system is one of the top applications of data science. Recommender systems are widely used today to suggest content to users to increase their engagement with a tool (e.g. watch another Netflix movie that you’ll probably like) or increase product sales (e.g. suggest relevant items you’ll probably buy). (Albini, 2023)

There are two main approaches to developing recommender systems:  
**content-based filtering**: based on features about products, movies, …(e.g. genres, actors, year, plot, or even calculated features using NLP), new content is suggested to the user.  
**collaborative filtering**: based on the user’s historical preferences, new content is suggested to the user.

**Problem Statement**

During the last few decades, with the rise of YouTube, Amazon, Netflix, and many other such web services, recommender systems have taken more and more place in our lives. From e-commerce (suggest to buyers articles that could interest them) to online advertisement (suggest to users the right contents, matching their preferences), recommender systems are today unavoidable in our daily online journeys. By analyzing the problems with ‘Book Recommendation System’ feature, how we can predict the best recommendation for users according to their items approach. A recommendation system helps an organization to create loyal customers and build trust by them desired products and services for which they came on your site. The recommendation system today is so powerful that they can handle the new customer too who has visited the site for the first time. They recommend the products which are currently trending or highly rated and they can also recommend the products which bring maximum profit to the company. In a very general way, recommender systems are algorithms aimed at suggesting relevant items to users (items being movies to watch, text to read, products to buy, or anything else depending on industries). Recommender systems are really critical in some industries as they can generate a huge amount of income when they are efficient or also be a way to stand out significantly from competitors. The main objective is to create a book recommendation system for users. We’ve all record and data with three different dataset – Book dataset (ISBN, Book-Title, Book-Author, Year-of Publication, Publisher, Image-URL-S, Image-URL-M, Image-URL-L); Users dataset (User-ID, Location, Age); Ratings dataset (User-ID, ISBN, Book-Rating). Providing specific data analysis and prediction to be done with this data. The main objective is to build a predictive recommender model, which could help in predicting – how we can predict the best recommendation for users according to their items approach. This would help us in providing better recommendation items to a right specific user.

• Book data – (ISBN, Book-Title, Book-Author, Year-Of-Publication, Publisher, Image-URL-S, ImageURL-M, Image-URL-L);   
• Users data - (User-ID, Location, Age);   
• Ratings data - (User-ID, ISBN, Book-Rating)

**Introduction**

A book recommendation system is a type of recommendation system where we have to recommend similar books to the reader based on his interest. The books recommendation system is used by online websites which provide ebooks like google play books, open library, good Read’s, Amazon Kindle etc. The major concern is about providing best recommendation to users. Looking at various factors/features we can take leverage of data to make better predictions on test set. To proceed with recommendation, it’s important to have all users interact with items. Generally, the models can predict the best recommendation for the data according to its RMSE score. This recommendation system data can show how it varies from every machine learning approach. So, generating different RMSE scores from different methods can reveal the best recommendation. Our goal is to build a predictive recommendation model using the SVD model, which could help companies in predicting get insights from user-items interactions and provide the best recommendation to users.

**Users-Items description of data**

The user’s interaction plays a very vital role for recommendation. To successfully build collaborative filtering model in recommender system, data preparation is important. Beginning with book data – dropping URL features (i.e. 'Image-URL-S', 'Image-URL-M', 'Image-URL-L'). We have some extra columns which are not required for our task like image URLs. And we rename the columns of each file as the name of the column contains space and lowercase letters so we will correct as to make it easy to use. The features depict great analysis by feature engineering.

**Observation from features and hypothetical assumption**

The dataset is reliable and can be considered as a large dataset. We have 271360 books data and total registered users on the website are approximately 278000 and they have given nearly 11 lakh rating. Hence, we can say that the dataset we have is nice and reliable. Agatha Christie is leading at the top with more than 600 counts, followed by William Shakespeare. It can happen in some possible cases that Agatha Christie is not the best Author, though Agatha Christie has the greatest number of books as compared to others. William Shakespeare is one of the popular Author in the world. Still, he doesn't have the highest number of books. Among all other Authors, it might happen that few of the Author might have some of the best seller books who have millions of copies been sold in world. Harlequin has the greatest number of books published, followed by Silhouette. Hypothetical assumptions - Some of the top Author's had published their books from Harlequin. We can observe Harlequin publisher's marking better performance than any other publishers. Penguin Books, Warner Books, Penguin USA, Berkely Publishing Group, and many more are among popular publisher's remarking competition with Harlequin. Though Penguin Books Publisher has a smaller number of books published but it might happen that only top Authors are approaching Penguin Books Publisher. There are 4618 entries as ‘0’ and 0 NaN entries in the Year of Publication field. Publication years are somewhat between 1950 - 2005.The publication of books got vital when it starts emerging from 1950. It might happen that people start to understand the importance of books and gradually get productivity habits in their life. Every user has their own taste in reading books based on what subject the Author uses. The subject of writing books got emerge from late 1940 slowly. Till 1970 it had the opportunity to recommend books to people or users what they love to read. The highest peak we can observe is between 1995-2001 year. The user understands what they like to read. Looking towards the raise the recommendation is also increased to understand their interest. Looking at the users age between 30- 40 prefer more and somewhat we can also view between 20-30. It is obvious that most of the user books are from Age 30 to 40. It might happen that the users are more interested in the subject of what Authors are publishing in the market. The age group between 20-30 are immensely attracted to read books published by Author. We can observe same pitch for Age group between 10-20 and 50-60. There can be a lot of different reasons. As per ratings "Selected Poems" has been rated most followed by "Little Women". Selected Poems are most favorable to users as per ratings. Three of the books 'The Secret Garden', 'Dracula', 'Adventures of Huckleberry Finn' are struggling to compete. Similarly, we can observe in 'Masquerade', 'Black Beauty', 'Frankenstein'. Firstly, the above ratings are unique ratings from 'ratings\_data' and 'books\_data' dataset. We have separated the explicit ratings represented by 1–10 and implicit ratings represented by 0. Let's make some hypothesis assumptions - Mostly the users have rated 8 ratings out of 10 as per books. It might happen that the feedback is positive but not extremely positive, as 10 ratings (i.e. best books ever). Now this count plot of book Rating indicates that higher ratings are more common amongst users and rating 8 has been rated highest number of times. There can be many assumptions based on ratings of users - taking ratings group from 1-4. This can have a negative impact for books been published if they have ratings from 1 to 4. It can be issues related to language, offended by any chapter's incident/paragraph/Author, they've read worst book ever. For 5 ratings the users might not be sure about book ratings whether it's positive or negative impact. take ratings group from 6-10. This is positive feedback - It can happen that not every book is perfect in all desire. So, the users have decided to rate 8. Since 6 ratings are very low among other ratings. As we can see, 7 and 8 are average and more ratings from users. 9 and 10 ratings are top best ratings based on Author's, Publisher's and Books been published.

**Top rated books as per ratings**

Building recommendation system based on popularity (i.e. ratings). These recommendations are usually given to every user irrespective of personal characterization. Merged book\_data dataset and ratings\_explicit. Considering ISBNs that were explicitly rated for this recommendation system. So we achieved top ten books as per ratings – book\_title book\_rating The Lovely Bones : A Novel 707 Wild Animus 581 The Da Vinci Code 494 The Secret Life of Bees 406 The Nanny Diaries : A Novel 393 The Red Tent (Bestselling Backlist) 383 Bridget Jones’s Diary 377 A Painted House 366 Life of Pie 336 Harry Potter and the Chamber of Secrets (Book 2) 326 The above are the top 10 books recommendation as per ratings. But this is not based on some recommendation system. They are top 10 books as per ratings.

**Steps Involved**

• Exploratory Data Analysis : Analytics for every dataset (i.e book, users, ratings) has helped to understand user-item interactions for book recommendation. Viewed top books as per ratings. Analysis based on top authors with highest number of books, top publishers with highest number of books, number of books published in yearly, users age distributions, top books as per ratings, different various user’s ratings.  
 • Null values treatment : We served with three dataset (i.e book\_dataset, users\_dataset, ratings\_dataset). We got null values in book dataset (features are book-author, publisher, Image-URL-L), users’ dataset in age. The ratings dataset doesn’t contain any null values.   
• Dropping and replacing data Proceeding with data cleaning and feature selection is a crucial step – we dropped feature like image URL. Replaced feature with lowercase and ‘-‘ to build a space between words. Some of the null values were present in feature data, we replaced with mean of that feature. Deal with mismatch features like book title, book\_author, year\_of\_publication, publisher. Considering age between 5- 90 we took users data to analysis and perform recommendation on it.  
 **Conclusion**SVD with RMSE score is 1.63 for this dataset. This performance could be due to various reasons : pattern of data, different models give different accuracy score, business understanding, machine learning approach etc

# Bibliography

Albini, G. (2023). Building a Movie Recommender Web App from Scratch with SVD and Flask: Part 1. *Towards Data Science* .