Capstone Project - 4

Book Recommendation System

By Shantanu Houzwala

Index

Title
❖ Introduction ❖ Problem Statement ❖ Understanding the Data ❖ Exploratory Data Analysis on Features ❖ Feature Engineering ❖ Models used to build Recommendation System ❖ Challenges faced ❖ Conclusion

Introduction

- 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.

- 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).

- Since, here we are trying to recommend books to users based popularity or ratings the user gave previously, I have tried different models like Popularity based recommender system and Collaborative filtering based recommender system.

Problem Statement

- 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 machine learning model to recommend relevant books to users based on popularity and user interests.

Understanding the Data

The Book-Crossing dataset comprises 3 files.

1. **Users**

Contains the users. User IDs (User-ID) have been anonymized and map to integers. Demographic data is provided (Location, Age) if available. Otherwise, these fields contain NULL values.

2. **Books**

Books are identified by their respective ISBN. Invalid ISBNs have already been removed from the dataset. Moreover, some content-based information is given (Book-Title, Book-Author, Year-Of-Publication, Publisher), obtained from Amazon Web Services. In the case of several authors, only the first is provided. URLs linking to cover images are also given, appearing in three different flavors (Image-URL-S, Image-URL-M, Image-URL-L), i.e., small, medium, large. These URLs point to the Amazon website.

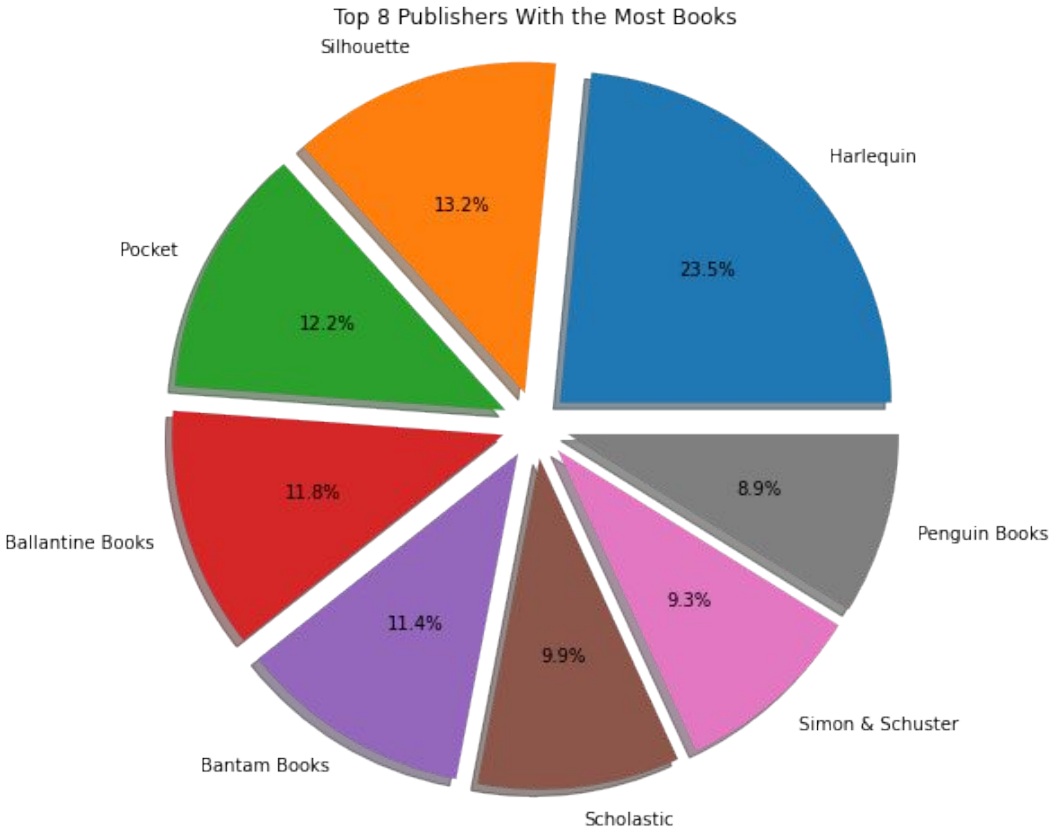
3. **Ratings**

Contains the book rating information. Ratings (Book-Rating) are either explicit, expressed on a scale from 1-10 (higher values denoting higher appreciation), or implicit, expressed by 0.

Exploratory Data Analysis

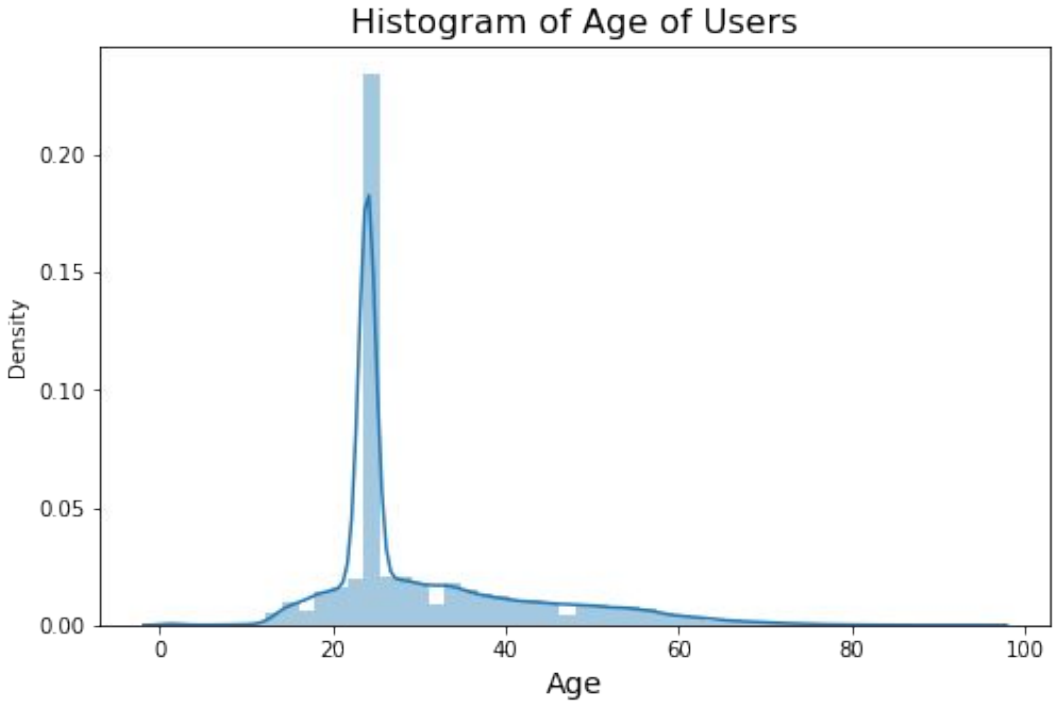
Finding Top 8 Publishers

- Finding Top 8 book publishers based on number of books published by publisher
- We can see that publisher Harlequin is the top publisher with 23.5% books published followed by Silhouette and Ballantine Books.



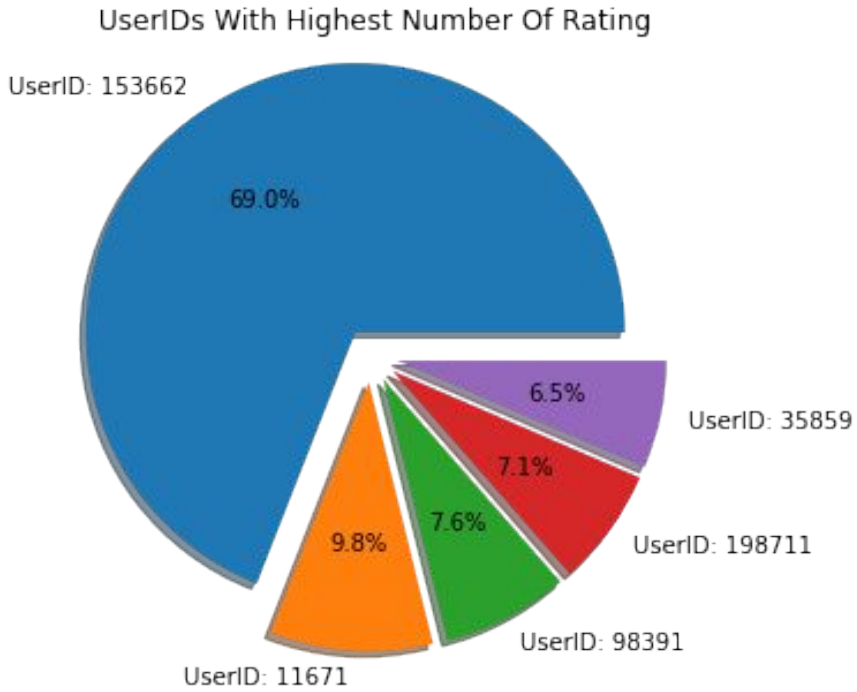
Distribution of Age of Users

- We can see that majority of the readers are of the age bracket 20–35
- This will help us understand the users better to recommend them books.



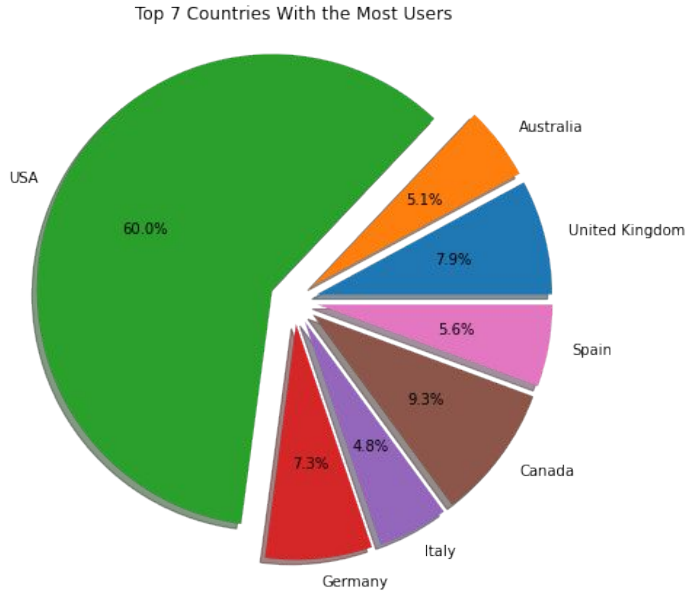
Users with highest ratings

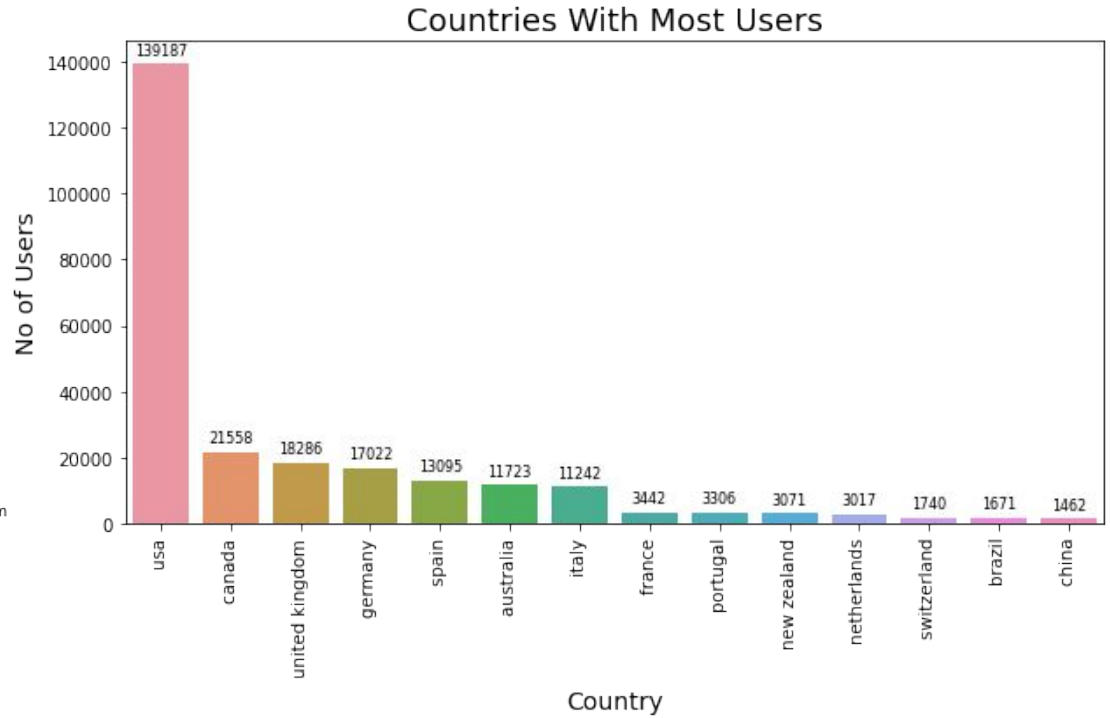
- We can see that UserID:153662 has given around 70% of the ratings out of 5 users.
- This information can be used to recommend books to similar users



Finding Countries with most number of Users

- From the bar chart and pie chart we can find that USA has most number of users followed by Canada, United Kingdom and Germany





FEATURE ENGINEERING

Feature Engineering on Year-Of-Publication

- From the analysis part we get that the Year-Of-Publication was wrongly mentioned for some of the rows.
- Diving deep into the Books dataframe we got to know that for these rows there was actually a column mismatch.

	ISBN	Book-Title	Book-Author	Year-Of-Publication	Publisher	Image-URL-M
209538	078946697X	DK Readers: Creating the X-Men, How It All Beg...	2000	DK Publishing Inc	http://images.amazon.com/images/P/078946697X.0...	http://images.amazon.com/images/P/078946697X.0...
220731	2070426769	Peuple du ciel, suivi de 'Les Bergers\";Jean-M...	2003	Gallimard	http://images.amazon.com/images/P/2070426769.0...	http://images.amazon.com/images/P/2070426769.0...
221678	0789466953	DK Readers: Creating the X-Men, How Comic Book...	2000	DK Publishing Inc	http://images.amazon.com/images/P/0789466953.0...	http://images.amazon.com/images/P/0789466953.0...

Feature Engineering on Year-Of-Publication

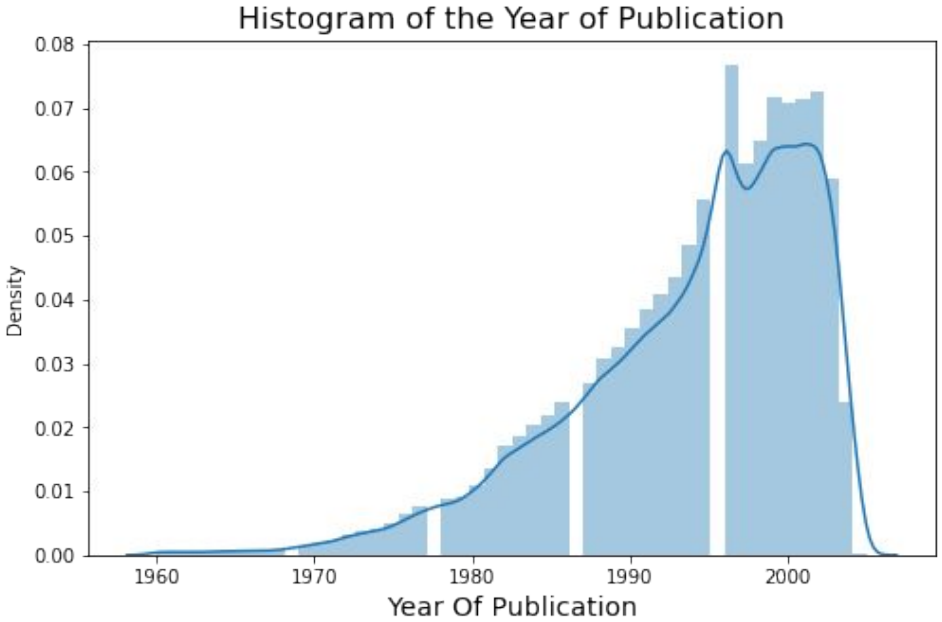
- As it can be seen from above table, there are some incorrect entries in the Year-Of-Publication field. It looks like Publisher names ‘DK Publishing Inc’ and ‘Gallimard’ have been incorrectly loaded as Year-Of-Publication in the dataset due to some errors in the csv file

- Making required corrections to 'Year of Publication' column

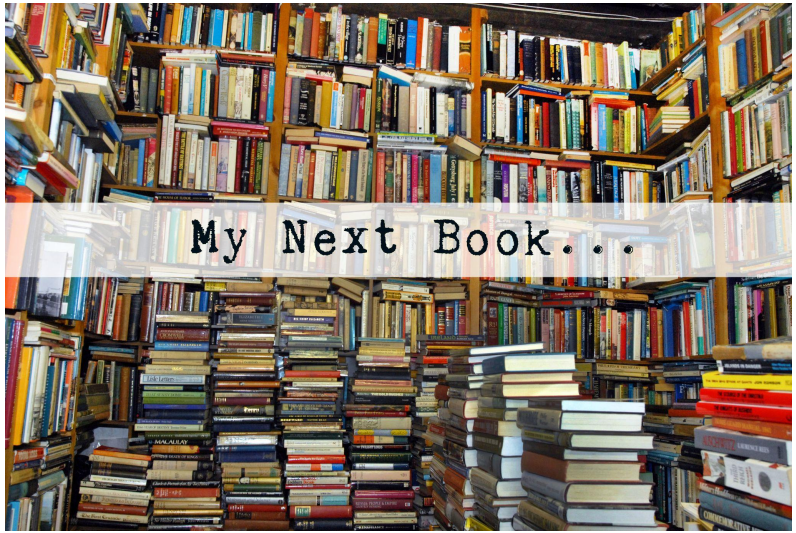
	ISBN	Book-Title	Book-Author	Year-Of-Publication	Publisher	Image-URL-M
209538	078946697X	DK Readers: Creating the X-Men, How It All Beg...	Michael Teitelbaum	2000	DK Publishing Inc	http://images.amazon.com/images/P/078946697X.0...
220731	2070426769	Peuple du ciel, suivi de 'Les Bergers'	Jean-Marie Gustave Le ClÃƒ?Ã‚Â©zio	2003	Gallimard	http://images.amazon.com/images/P/2070426769.0...
221678	0789466953	DK Readers: Creating the X-Men, How Comic Book...	James Buckley	2000	DK Publishing Inc	http://images.amazon.com/images/P/0789466953.0...

Feature Engineering on Year-Of-Publication

- For Year-Of-Publication we observed that the year mentioned was beyond 2020 for some entries whereas the dataset was created in 2004.
- For the anomalous entries we first fill them with Nan values. Replacing NaN values with median values
- Plotting distribution of Year of Publication from 1960 to 2005



Building Recommender System



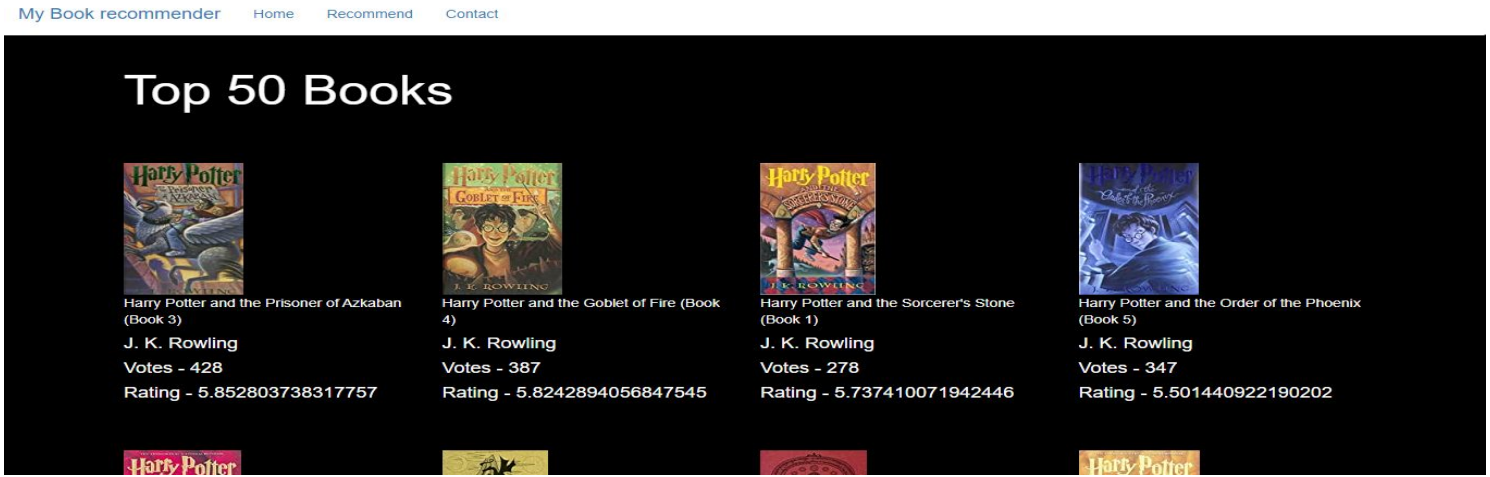
Models used to build Recommendation System

- Popularity Based Approach
- Collaborative Filtering

Popularity Based Approach

- It is a type of recommendation system which works on the principle of popularity and or anything which is in trend. These systems check about the books which are in trend or are most popular among the users and directly recommend them.

- I have sorted top 50 books from the dataset on the basis of number of ratings received(more than 250) and highest average rating.

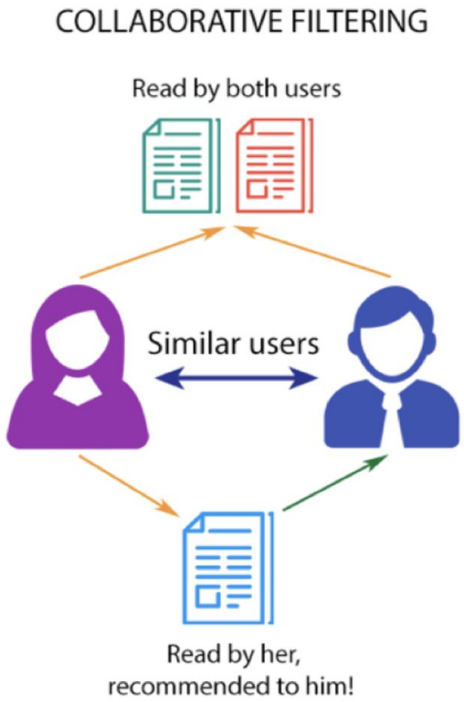


Collaborative Filtering

AI

- Collaborative Filtering is considered to be one of the very smart recommender systems that work on the similarity between different users and also items that are widely used as an e-commerce website and also online movie websites. It checks about the taste of similar users and makes recommendations.

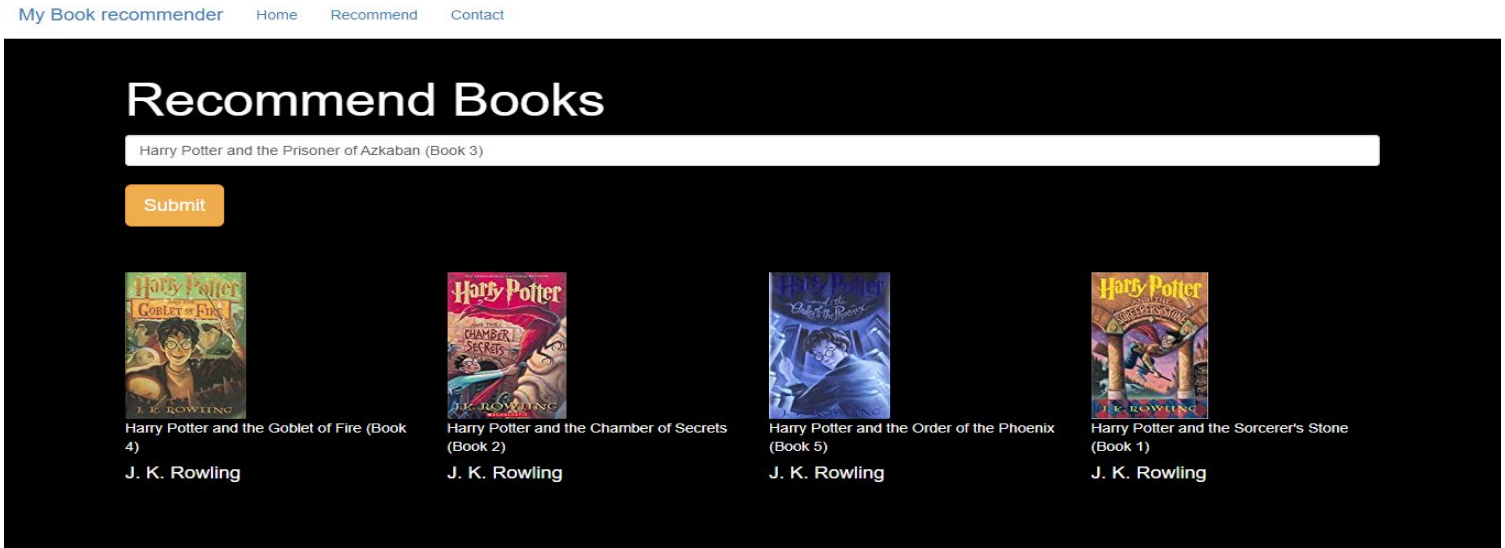
- The similarity is not restricted to the taste of the user, moreover there can be consideration of similarity between different items also. The system will give more efficient recommendations if we have a large volume of information about users and items.



Collaborative Filtering

- For this model, with the help of Cosine Similarity measurement, I have created the correlation matrix considering only those books which have total ratings of more than 50 and users who have rated more than 200 books.

- For the input book using the correlation matrix, top 4 books are recommended.



Challenges Faced

- Handling of sparsity was a major challenge since the user interactions were not present for the majority of the books.
- Understanding the metric for evaluation was a challenge as well.
- Since the data consisted of text data, data cleaning was a major challenge in features like Year-Of-Publication.
- Decision making on missing value imputations.

Conclusion

- In EDA, we found that majority of the readers were of the age bracket 20–35 and most of them came from North American and European countries namely USA, Canada, UK, Germany and Spain.

- Top rated books were essentially novels. Books like Harry Potter series, Lord of the Rings series, The Lovely Bone, etc.

- The objective of this project to build a book recommendation system is achieved and the model works well.

❖ **FUTURE SCOPE**

- Given more information regarding the books dataset, namely features like Genre, Description etc., we could implement a content-filtering based recommendation system and compare the results with the existing collaborative-filtering based system.

- We would like to explore various clustering approaches for clustering the users based on Age, Location etc., and then implement voting algorithms to recommend items to the user depending on the cluster into which it belongs.

Thank You