Capstone Project - 4

Book Recommendation System

By – Rahul Shah

CONTENTS OF THE PRESENTATION

- Problem Statement
- Data Overview
- Data Cleaning
- Exploratory Data Analysis
- Model Creation
- Model Evaluation
- Challenges
- Conclusion

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.

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.

Data Overview

- Understanding datasets better:

The Book-Crossing dataset comprises 3 files.

Users

Contains the users. Note that 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.

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. Note that 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.

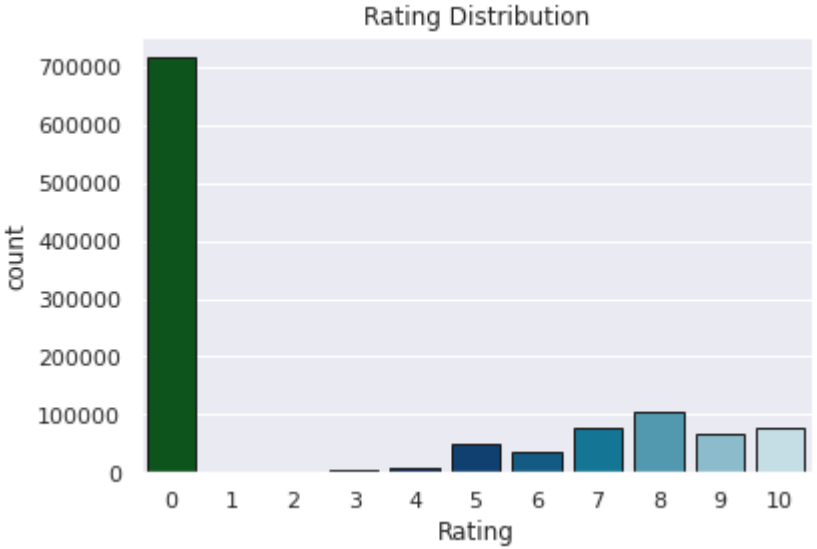
Ratings

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.

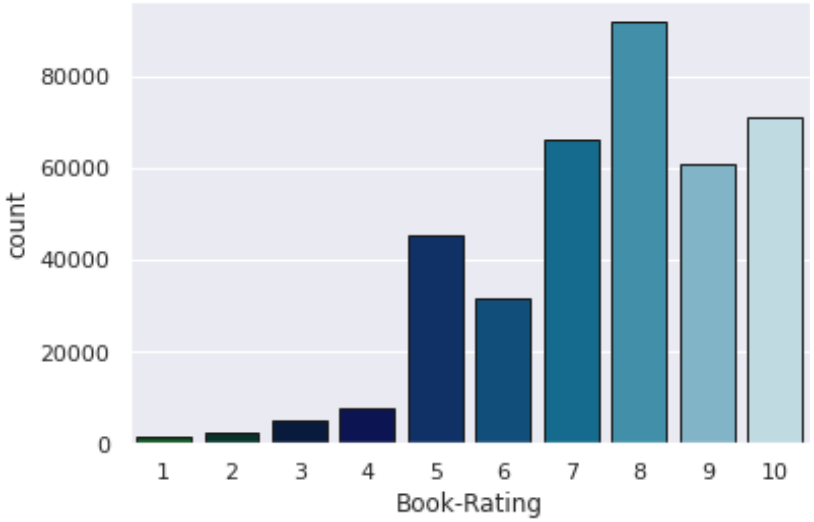
Data Cleaning

- In books data, ‘Book-Author’ and ‘Publisher’ column had missing values, hence replaced these null values by searching over internet.
- Age column had around 40% of missing values, hence replaced these null values with median values based on the country.
- Ratings data has no missing values
- In ‘Book-Author’, ‘Book-Title’ and ‘Year-Of-Publication’ columns there were some incorrect entries, since the entries were few so manually corrected these columns.

Exploratory Data Analysis



a. Number of book without rating

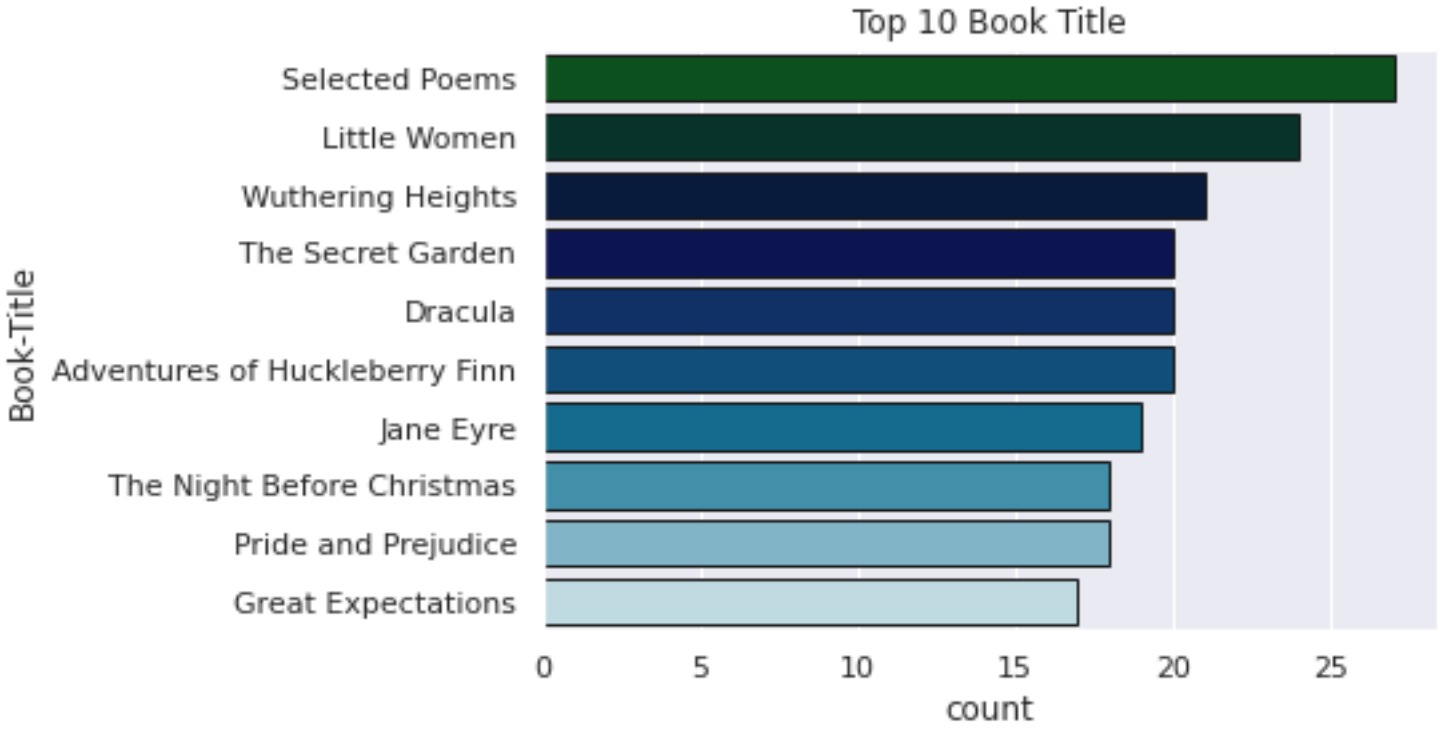


b. Removing the zero values from rating

higher ratings are more common amongst users and rating 8 has been rated highest number of times.

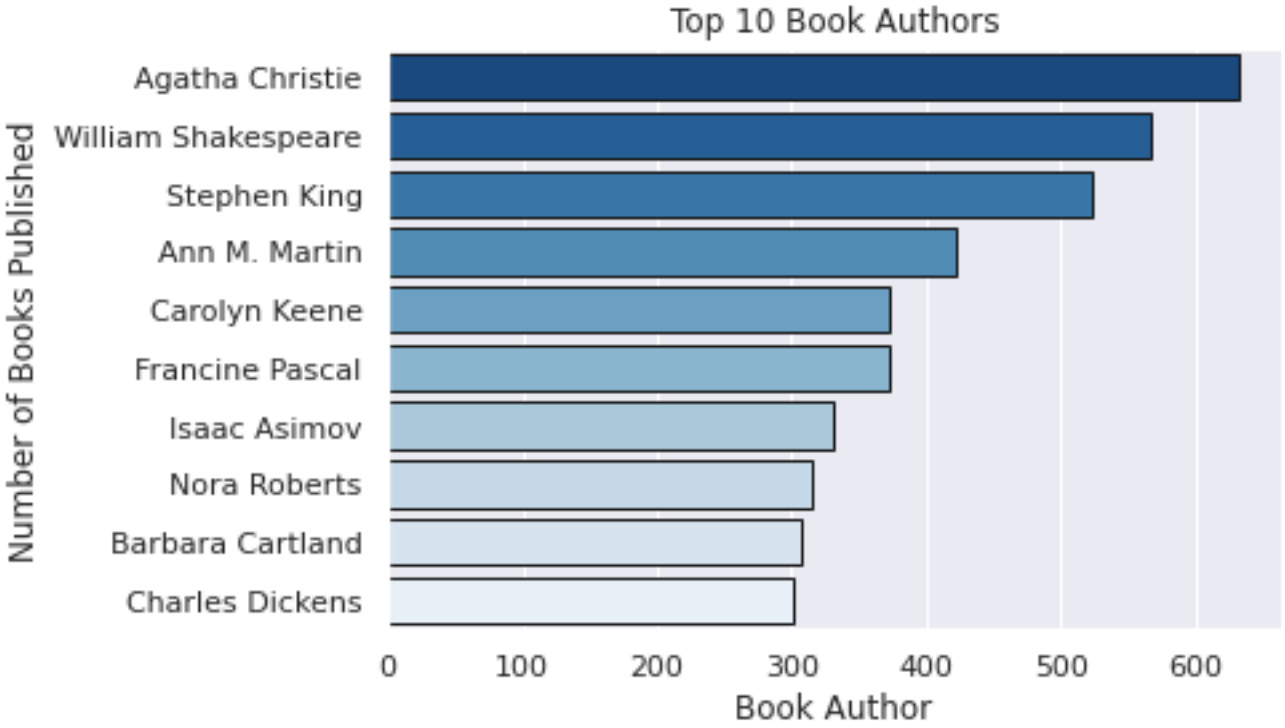
EDA continued…

Top most Book Title



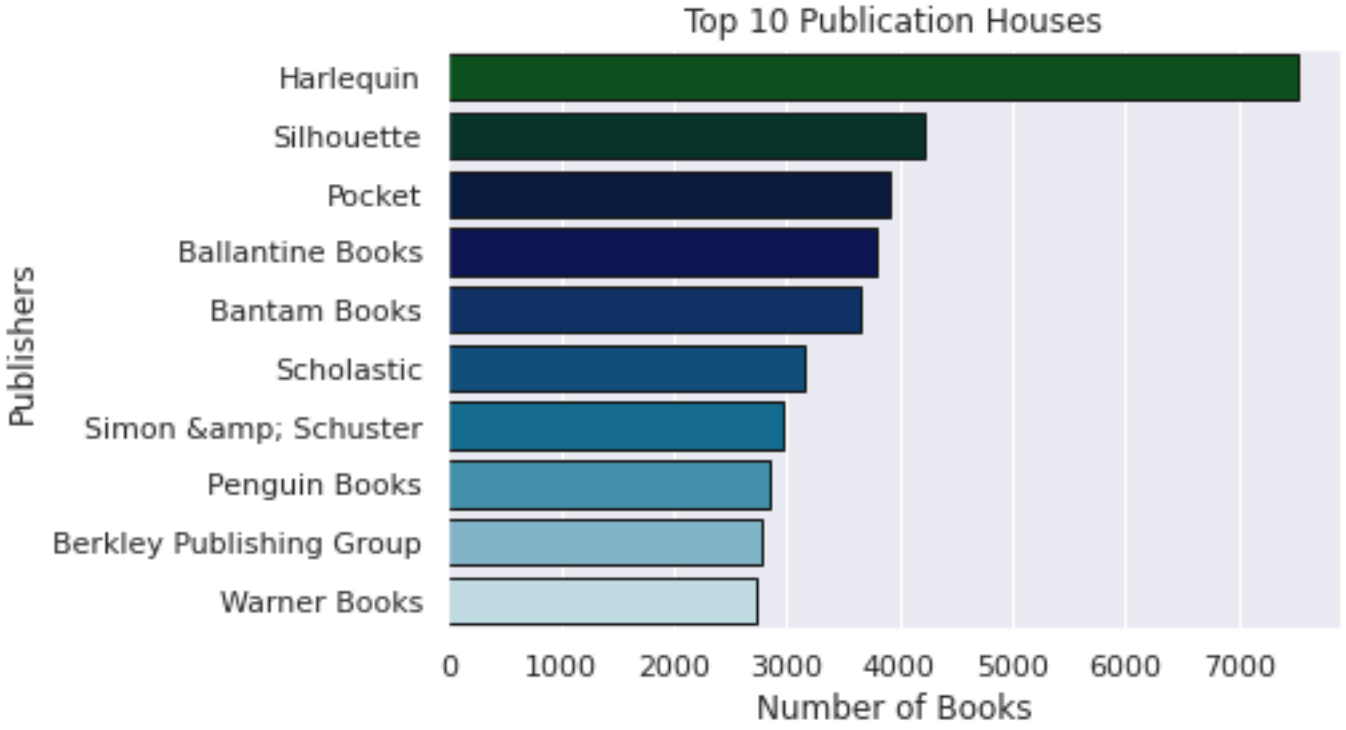
EDA continued…

Agatha Christie wrote highest number of books in our given dataset.



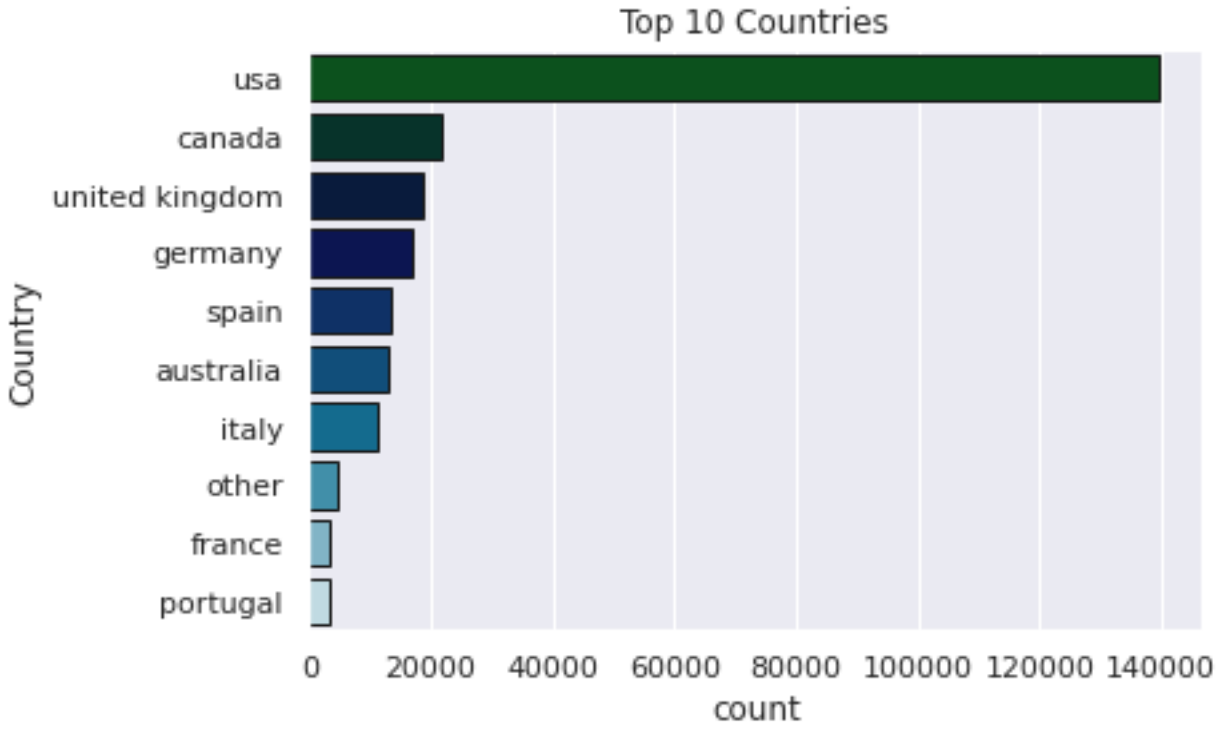
EDA continued…

Harlequin published highest number of books in our given dataset.



EDA continued…

USA has the highest number of users.



Modelling

1.)Popularity Based Recommendation

Book weighted average formula:

Weighted Rating(WR)=[vR/(v+m)]+[mC/(v+m)]

Where,

v is the number of votes for the books;
m is the minimum votes required to be listed in the chart;
R is the average rating of the book; and
C is the mean vote across the whole report.

Popularity based recommended books

	Book-Title	Book-Author	Total_Users_Rated	Average_Rating	Score
0	Harry Potter and the Goblet of Fire (Book 4)	J. K. Rowling	137	9.262774	8.741835
1	Harry Potter and the Sorcerer's Stone (Harry P...	J. K. Rowling	313	8.939297	8.716469
2	Harry Potter and the Order of the Phoenix (Boo...	J. K. Rowling	206	9.033981	8.700403
3	To Kill a Mockingbird	Harper Lee	214	8.943925	8.640679
4	Harry Potter and the Prisoner of Azkaban (Book 3)	J. K. Rowling	133	9.082707	8.609690
5	The Return of the King (The Lord of the Rings,...	J.R.R. TOLKIEN	77	9.402597	8.596517
6	Harry Potter and the Prisoner of Azkaban (Book 3)	J. K. Rowling	141	9.035461	8.595653
7	Harry Potter and the Sorcerer's Stone (Book 1)	J. K. Rowling	119	8.983193	8.508791
8	Harry Potter and the Chamber of Secrets (Book 2)	J. K. Rowling	189	8.783069	8.490549
9	Harry Potter and the Chamber of Secrets (Book 2)	J. K. Rowling	126	8.920635	8.484783

Collaborative Filtering (CF)

- The most prominent approach to generate recommendations.
  - Used by large, commercial e-commerce sites
  - Well understood, various algorithms and variations exist
  - Applicable in many domains (books, movies)

- Approach
  - Use the “wisdom of the crowd” to recommend items

Data Preparation for kNN

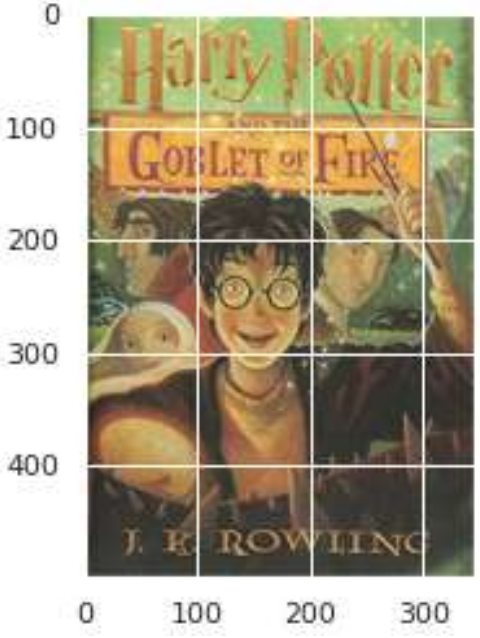
- Extracted 50 of users and ratings.
- Extracted books that have received more than 50 ratings.
- Merged ratings data and books data.
- Created a pivot table where columns are user id and indexes are books and values are ratings

Collaborative Filtering using k-Nearest Neighbors (KNN)

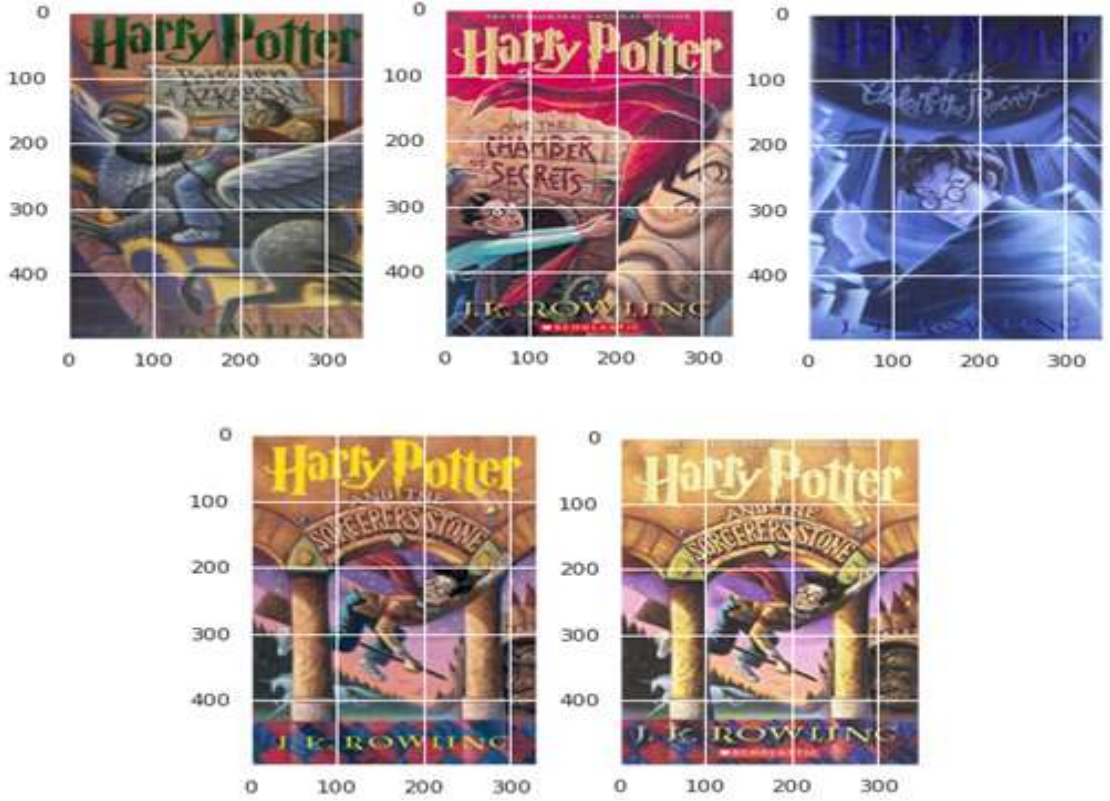
This filtering technique uses the similarity between books using the ratings given by the users. The underlying assumptions of this technique is that ‘**a reader gives similar ratings to a similar books**’ This approach uses NearestNeighbors from the sklearn library. To keep up with the computing efficiency of the system, only books with atleast 50 ratings and users who have rated at least 50 books are considered.

Result of k-NN

Target Book



Recommendations



Data Preparation for SVD

- Consider only those ratings that are not equal to zero.
- Filter ISBN and users with at least 10 interactions.

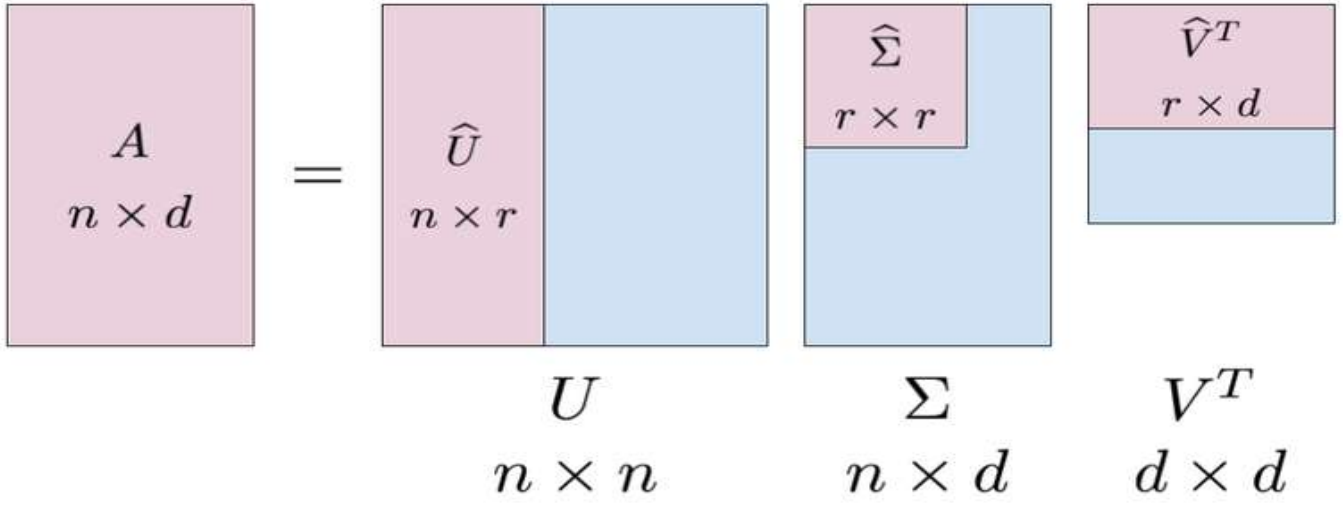
Observations:

- After dropping 0 ratings, remaining rows are 41419.
- After filtering out ISBN and Users with at least 10 interactions, remaining rows are 37846.

Collaborative Filtering using SVD

Singular Value Decomposition is a way of breaking up (factoring) a matrix into three simpler matrices.

A = U Σ VT



Evaluation

In Recommender Systems, there are a set metrics commonly used for evaluation. We choose to work with **Top-N accuracy metrics**, which evaluates the accuracy of the top recommendations provided to a user, comparing to the items the user has actually interacted in test set.

This evaluation method works as follows:

- For each user
  - For each item the user has interacted in test set
    - Sample 100 other items the user has never interacted.
    - Ask the recommender model to produce a ranked list of recommended items, from a set composed of one interacted item and the 100 non-interacted items
    - Compute the Top-N accuracy metrics for this user and interacted item from the recommendations ranked list
- Aggregate the global Top-N accuracy metrics

Evaluation Contd.

```
Evaluating Collaborative Filtering (SVD Matrix Factorization) model...
1306 users processed

Global metrics:
{'modelName': 'Collaborative Filtering', 'recall@5': 0.31175693527080584, 'recall@10': 0.4355350066050198, 'recall@15': 0.51889035667107}
```

	hits@5_count	hits@10_count	hits@15_count	interacted_count	recall@5	recall@10	recall@15	User-ID
0	17	24	26	59	0.29	0.41	0.44	16795
16	13	19	26	58	0.22	0.33	0.45	98391
47	19	23	26	54	0.35	0.43	0.48	153662
7	30	36	43	52	0.58	0.69	0.83	114368
128	23	28	32	51	0.45	0.55	0.63	104636
159	6	7	12	47	0.13	0.15	0.26	95359
336	9	14	17	41	0.22	0.34	0.41	158295
59	21	29	29	34	0.62	0.85	0.85	123883
347	8	10	15	32	0.25	0.31	0.47	60244
109	5	7	9	30	0.17	0.23	0.30	35859

Challenges

- High Volume of Data.
- Understanding the metric for evaluation.
- Crashing of Session due to large pivot matrix.
- Missing value imputation and outlier treatment was also quite challenging.

Conclusion

Throughout the study, we performed various steps to build a book recommender system. We started with data wrangling in which we tried to handle null values and performed feature modifications. Next, we did some exploratory data analysis and tried to draw observations out of it.

Finally, A book recommendation system was designed using different filtering techniques. After implementing Collaborative Filtering model (SVD matrix factorization), we are satisfied with the results. We observe that we got Recall@10 (43%) and Recall@15 (52%), which is fair enough for such a large dataset.

Conclusion from analysis includes:

- "Selected Poems" were read more by the users.
- Majority of the users are from USA.
- "Harlequin" has published the most number of books.
- Among Authors, "Agatha Christie" has written the most number of books followed by "William Shakespeare" and "Stephen King".

THANK YOU