Capstone Project – 4

Book Recommendation System

By – Deeshu Kr. Pandit

Data Science trainee, Almabetter

Agenda

1. Introduction and Problem Description
2. Data Description
3. Feature engineering and Exploratory data analysis
4. Recommendation Systems
5. Evaluation
6. Summary and conclusions

1. Problem Description

- 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 buyer 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 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.

- **Business Goal**
- We are required to do basic EDA on the given data for the Amazon web services and analyze some insights from it. Finally, we have to use recommendation system to recommend the unexplored books to the users.

2. Data Description

- **The Book-Crossing dataset comprises 3 files –**

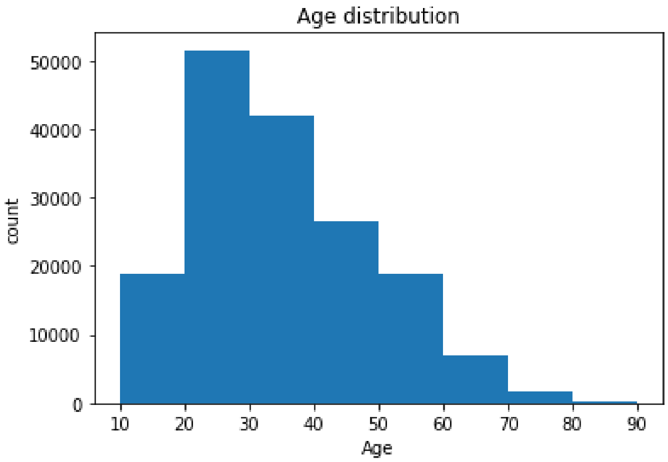
- **Users:** Contains the users. Note that user IDs (User-ID) have been anonymized and map to integers. Demographic data is provided (Location, Age) if available. Otherwise, these fields contain NULL values.

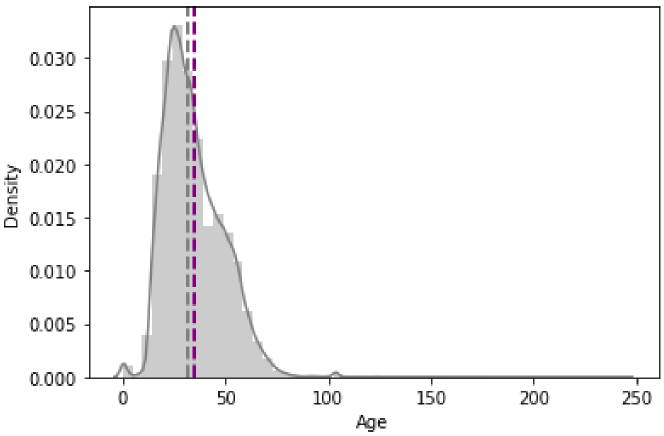
- **Books:** Books are identified by their respective ISBN. Invalid ISBNs have already been removed from the dataset. Moreover, some content-based information is given (Book-Title, Book-Author, Year-Of-Publication, Publisher), obtained from Amazon Web Services. Note that in the case of several authors, 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 explicit, expressed on a scale from 1-10 (higher values denoting higher appreciation).

3. EDA and Feature Engineering

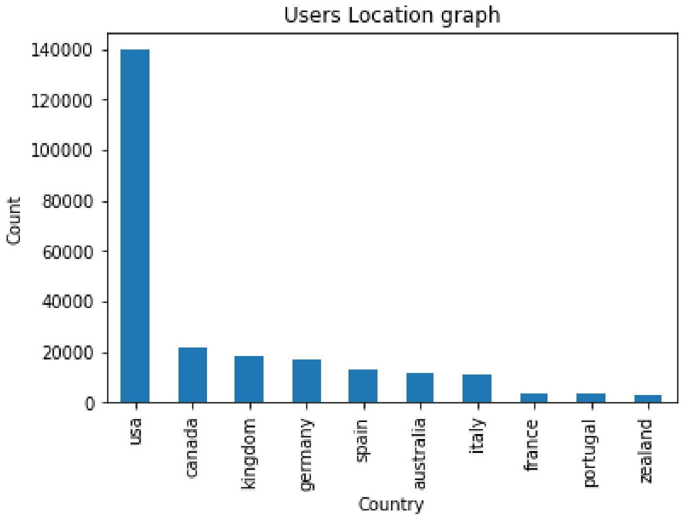
Age distribution of Users in the dataset



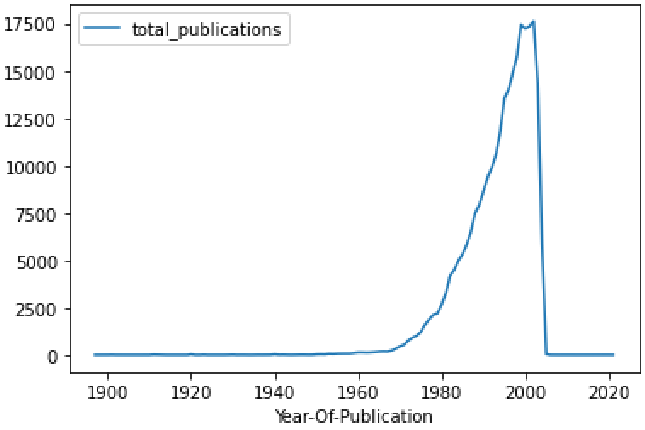


EDA and Feature Engineering (cont..)

- Derived new feature “Country” from “Location” feature to understand the distribution of location of users.

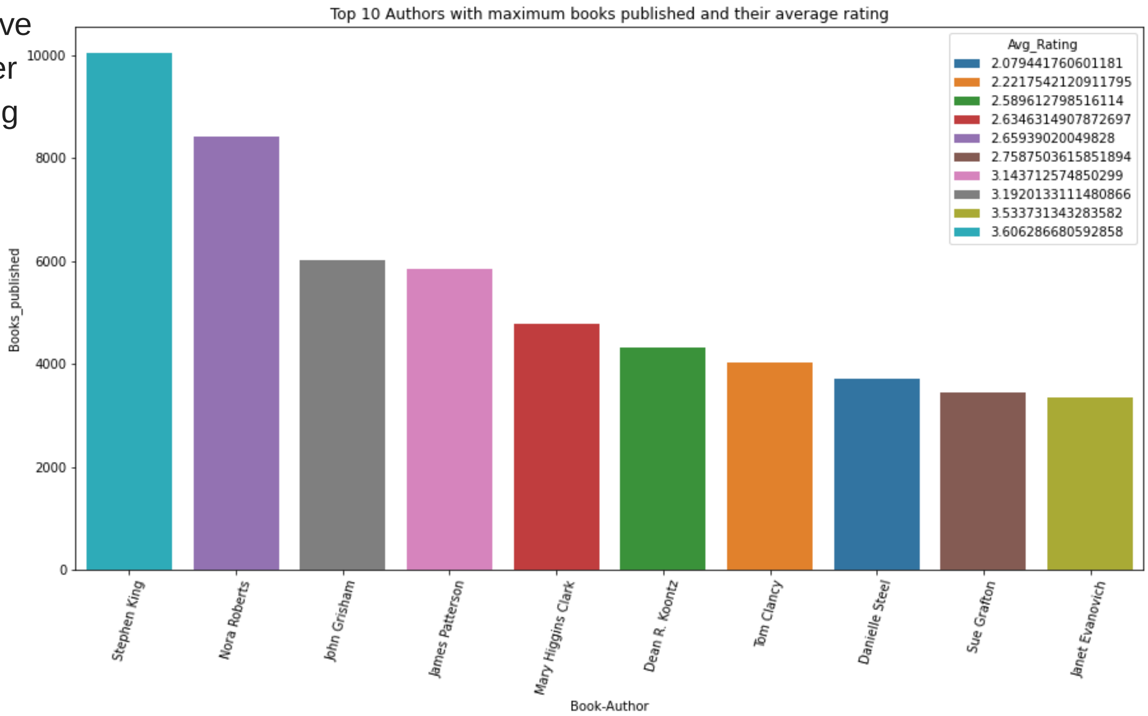


- Analyzing the trend in the publication of books per year.



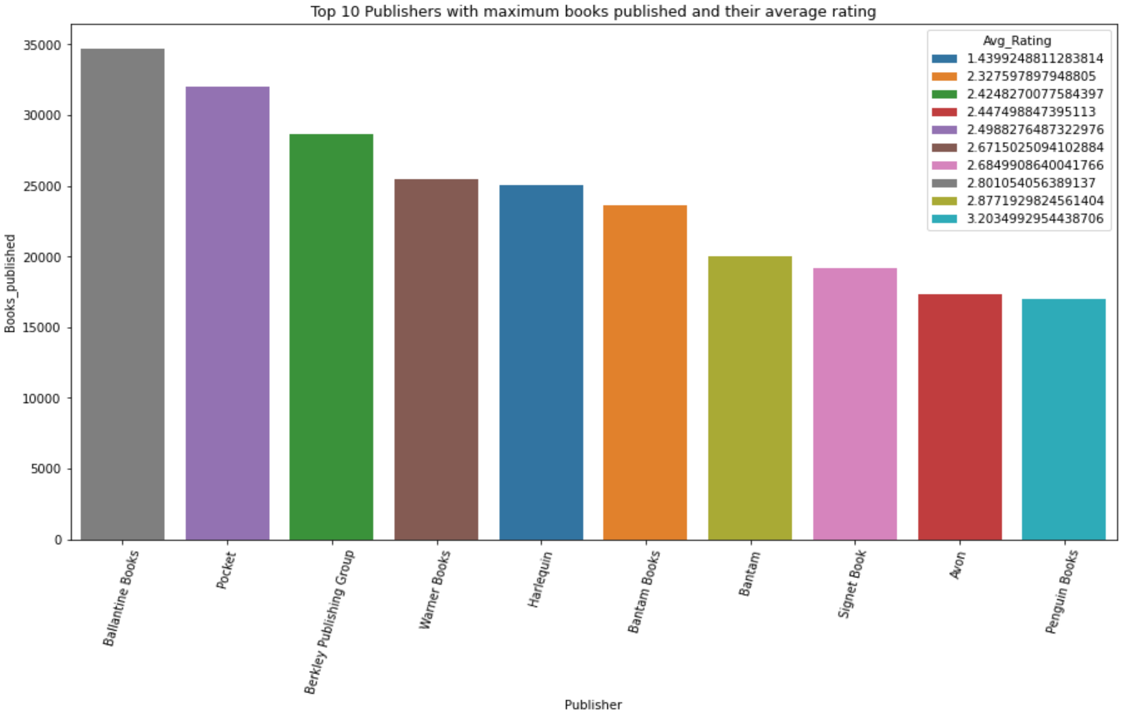
EDA and Feature Engineering (cont..)

- The top 10 Authors who have published maximum number of books and their avg rating of all books.



EDA and Feature Engineering (cont..)

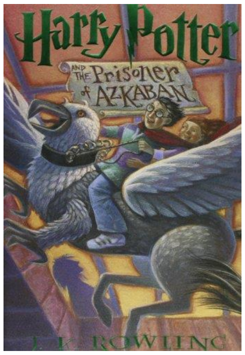
- The top 10 Publisher who have published maximum number of books and their avg rating of all books.



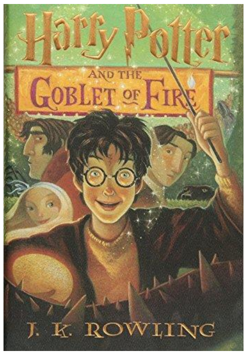
4. Recommendation Systems

❑ Popularity based Recommendations

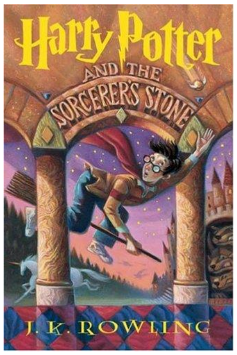
- Here we tried to find the top 20 books published yet based on the average ratings of all users, which can be recommended to all users as a general recommendation on home page of the platform.
- The books we have considered are the only which has been rated by more than 250 users.
- The top 5 books amongst the top 20 books are listed below -



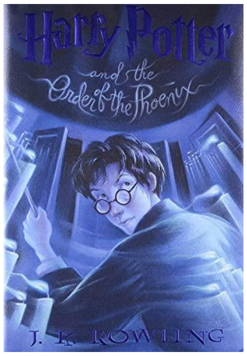
Total_ratings	Avg_rating
428	5.85



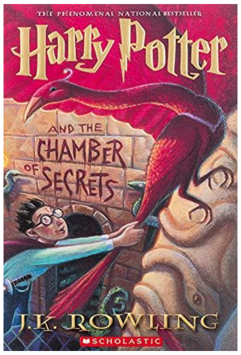
Total_ratings	Avg_rating
387	5.82



Total_ratings	Avg_rating
278	5.74



Total_ratings	Avg_rating
347	5.50



Total_ratings	Avg_rating
556	5.18

Recommendation Systems(cont..)

❑ Collaborative Filtering Based Recommender System

- Collaborative filtering is a technique that can filter out items that a user might like based on reactions by similar users
- It works by searching a large group of people and finding a smaller set of users with tastes similar to a particular user. It looks at the items they like and combines them to create a ranked list of suggestions.

Memory Based Approach

- Statistical techniques are applied to the entire dataset to calculate the predictions.
- To find the rating **R** that a user **U** would give to an item **I**, the approach includes:
    1. Finding users similar to **U** who have rated the item **I.**
    2. Calculating the rating **R** based on the ratings of users found in the previous step.

Model Based Approach

- It involve a step to reduce or compress the large but sparse user-item matrix.
- Matrix factorization is used as breaking down a large matrix into a product of smaller ones.

Recommendation Systems(cont..)

❑ Memory Based Approach(Item-based)

- Built a user-item pivot matrix , which contains the ratings given by users to the books.
- Calculated the similarity score to compare the similarity between the items.
- Defined a function **recommend** which will take an item as an input which is interacted by user in the past and based on the similarity score new items will be recommended to the user which he has not interacted yet.

```
recommend('The Notebook’)
```

```
A Walk to Remember
The Rescue
One Door Away from Heaven
Toxin
The Five People You Meet in Heaven
```

Recommendation Systems(cont..)

❑ Model Based Approach

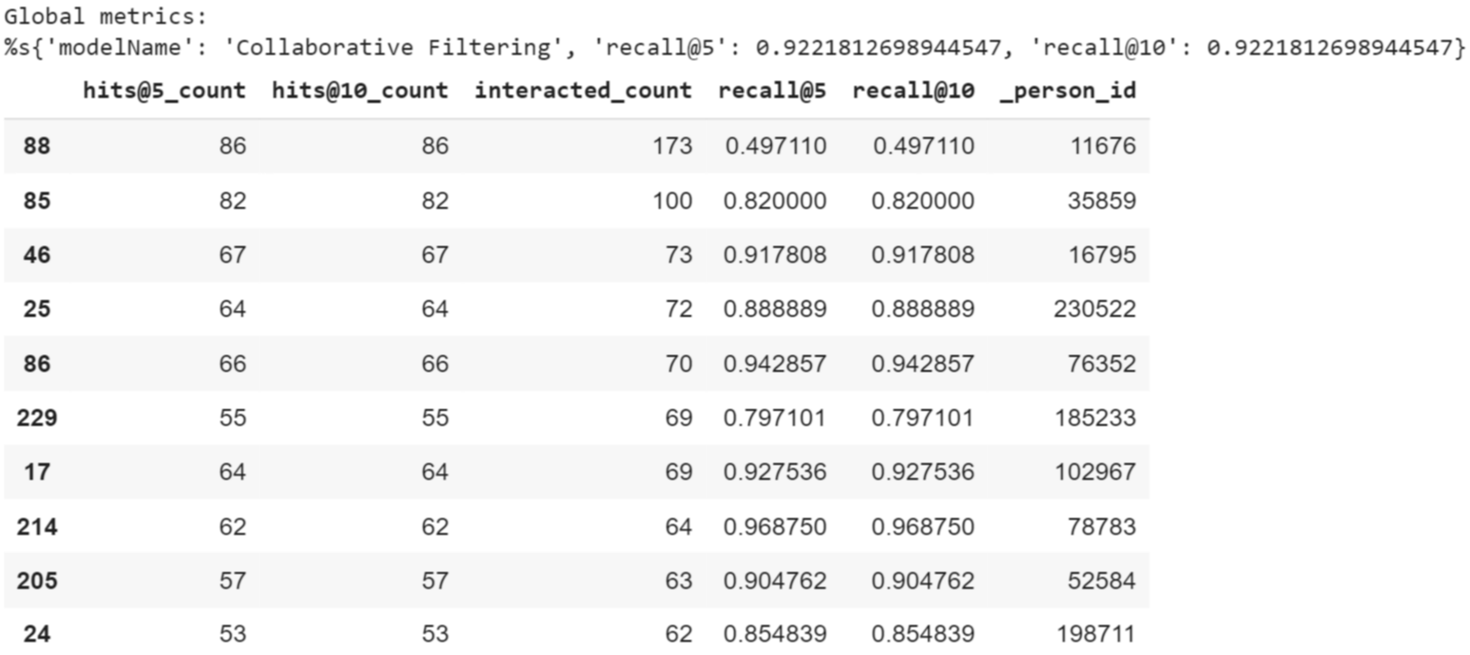
- Built a user-item pivot matrix , which contains the ratings given by users to the books.
- Used Singular Value Decomposition (SVD) to break down a large matrix into a product of smaller ones.
- Reconstructed the matrices to get original one.
- Finally built a function to compare the reconstructed matrix and original matrix to get the rating values of the items which has not been rated yet.

Model 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.

Recommendation Systems(cont..)

Model Evaluation



```
Global metrics:
%s{'modelName': 'Collaborative Filtering', 'recall@5': 0.9221812698944547, 'recall@10': 0.9221812698944547}
```

	hits@5_count	hits@10_count	interacted_count	recall@5	recall@10	_person_id
88	86	86	173	0.497110	0.497110	11676
85	82	82	100	0.820000	0.820000	35859
46	67	67	73	0.917808	0.917808	16795
25	64	64	72	0.888889	0.888889	230522
86	66	66	70	0.942857	0.942857	76352
229	55	55	69	0.797101	0.797101	185233
17	64	64	69	0.927536	0.927536	102967
214	62	62	64	0.968750	0.968750	78783
205	57	57	63	0.904762	0.904762	52584
24	53	53	62	0.854839	0.854839	198711

Summary and conclusions

1. After importing and analyzing the data, following insights were observed-

•The age distribution of users are approximately normal having median age as 32.
•The top 5 countries from where the users belong are - USA, Canada, Kingdom, Germany and Spain, followed by Australia and Italy.
•There is an exponential growth in the publication of books from 1970 to 2000.
•The top 5 Authors having most book published are Stephen King, Nora Roberts, John Grisham, James Patterson and Mary Higgins Clark.
•The top 5 Publishers having most book published are Ballantine Books, Pocket, Berkley Publishing Group, Warner Books and Harlequin.

2. The recommendation Systems we used are –

•Popularity based recommendations
•Collaborative filtering recommendation system -

1. Memory based approach
2. Model based approach

•The model-based approach got **Recall@5** = (92%) and also **Recall@10** = (92%) .

Thank You