BOOK RECOMMENDATION SYSTEM

Krushnagopal Brahma
Data science trainee,
Alma Better, Bangalore

Abstract:

Recommender systems are machine learning systems that help users discover new product and services. 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. A book recommendation system is a type of recommendation system where we have to recommend similar books to the reader based on his/her interest. The books recommendation system is used by online websites which provide ebooks like google play books, open library, goodReads, etc.

1. 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). Recommendation 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.

2. Dataset 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.

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.

Books Data :

- Data set Contains 'ISBN', 'Book-Title', 'Book-Author', 'Year-Of-Publication', 'Publisher', 'Image-URL-S', 'Image-URL-M', 'Image-URL-L‘ information's.
- The Data set Contains 271354 Rows

Users Data :

- Data set Contains ‘User-ID’ , ‘Location’, ‘Age’ information's.
- The Data set Contains 278858 Rows

Ratings Data :

- Data set Contains ‘User-ID’ , ‘ISBN’, ‘Book-Ratings’ information's.
- The Data set Contains 1149780 Rows

3. Data Preprocessing

- we drop those column which contain url of book image because we work on book recommendation no need image and its urls. So we drop 3 columns which contain image urls.
- Without removing nan valu we take nan valu as other catagory
- The publication must be int format but there are some error in it at some phase it show yr='DK Publishing Inc', yr='Gallimard' its not posible .
- There are 11072 valu are null we can't remove these many values so we try to cover it with mean age value.
- we se in our user data less than 1% data is bellow 13 yr but min age is 0 and almost 99% data cover on age of 71 but max age is 244.So we consider the users age which is less than 10 and higher than 75 we consider it as out layer .We cap our out layer with mean value.

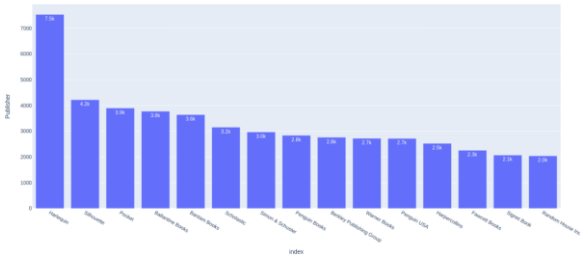
- we se user id 11676 gives rating on 13602 books rating its consider as most impact full user. Theare are so many user who are only given rating on 1 book so they might nt proper compression rating to our ratting we consider if some user gives rating in more than 200 books he might be good impact full rating user.

4. EDA :

- After mounting our drive and fetching and reading the dataset given, we performed the Exploratory Data Analysis for it.
- To get the understanding of the data and how the content is distributed in the dataset, its type and details such as which countries are watching more and which type of content is in demand etc. has been analyzed in this step.

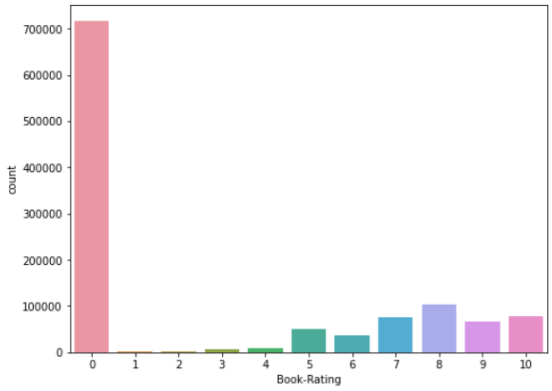
Index vs Book-Author:

we se that Harlequin publicatation publish most no of book its almost double than Silhouette which is in 2nd place.

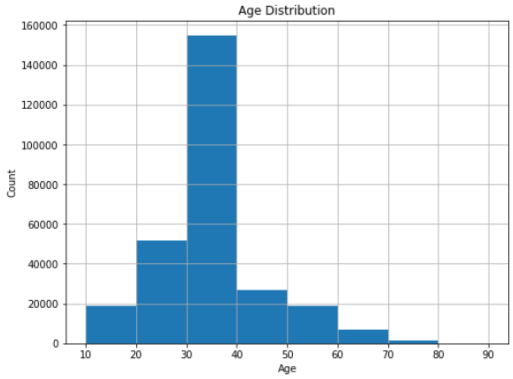


Count vs Book-Rating:

- If we se the rating part most of not given any review here we consider 0.
- It almost 50 to 60 percent of our dataset.



Count vs Age:

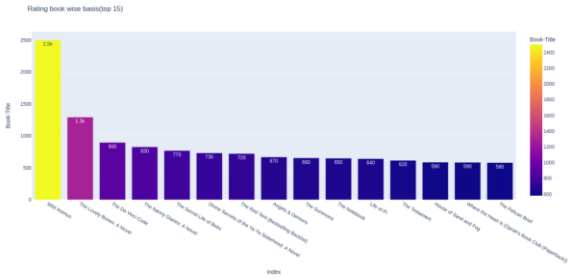


- Most of the user who rated the mv are from 30-40 yrs age then in 2nd its less than half of that catagory is 20-30.

Book Time Vs Index:

In book Wild Animus user gives a high no of rating which is 2.5k.

- In 2nd place The Lovely Bones:A Novel book with rating of 1.3k.
- Except these 2 books all other books are rating in between 580 to 900.



3. Clustering

Clustering (also called cluster analysis) is a task of grouping similar instances into clusters. More formally, clustering is the task of grouping the population of unlabeled data points into clusters in a way that data points in the same cluster are more similar to each other than to data points in other clusters. The clustering task is probably the most important in unsupervised learning, since it has many applications, for example:

- data analysis: often a huge dataset contains several large clusters, analyzing which separately, you can come to interesting insights.
- anomaly detection: as we saw before, data points located in the regions of low density can be considered as anomalies
- semi-supervised learning: clustering approaches often helps you to automatically label partially labeled data for classification tasks.
- Indirectly clustering tasks (tasks where clustering helps to gain good results): recommender systems, search engines, etc.
- directly clustering tasks: customer

segmentation, image segmentation, etc .

4. Book Recommendation System

1. Popularity Based (Books reviewed by most of the user)

Top 5 Popular books are:

	ISBN	Book-Title	Book-Author	Year-Of-Publication	Publisher	No_of_review
1	0316666343	The Lovely Bones: A Novel	Alice Sebold	2002	Little, Brown	707
0	0971880107	Wild Animus	Rich Shapero	2004	Too Far	581
3	0385504209	The Da Vinci Code	Dan Brown	2003	Doubleday	487
2	0312195516	The Red Tent (Bestselling Backlist)	Anita Diamant	1998	Picador USA	383
4	0060928336	Divine Secrets of the Ya-Ya Sisterhood: A Novel	Rebecca Wells	1997	Perennial	320

2.Popular books in country wise

```
Enter the name of place: india
india
```

	ISBN	Book-Title	Book-Author	Year-Of-Publication	Publisher	No_of_review
0	0971880107	Wild Animus	Rich Shapero	2004	Too Far	3
1	0486284735	Pride and Prejudice (Dover Thrift Editions)	Jane Austen	1995	Dover Publications	2
2	0671047612	Skin And Bones	Franklin W. Dixon	2000	Aladdin	2
3	8171670407	Inscrutable Americans	Mathur Anurag	1996	South Asia Books	2
4	0446357421	If Tomorrow Comes	Sidney Sheldon	1988	Warner Books	1

3. Books popular yr wise

	ISBN	Book-Title	Book-Author	Year-Of-Publication_x	Publisher	Year-Of-Publication_y	no_of_user_review
0	0002005018	Clara Callan	Richard Bruce Wright	2001	HarperFlamingo Canada	2001	9
1	0060973129	Decision in Normandy	Carlo D'Este	1991	HarperPerennial	1991	2
2	0374157065	Flu: The Story of the Great Influenza Pandemic...	Gina Bari Kolata	1999	Farrar Straus Giroux	1999	6
3	0399135782	The Kitchen God's Wife	Amy Tan	1991	Putnam Pub Group	1991	17
4	0425176428	What If?: The World's Foremost Military Histor...	Robert Cowley	2000	Berkley Publishing Group	2000	1
...	...	...	...	...	...	...	...
137707	0385264707	Dreamsnake	Vonda N. McIntyre	1978	Houghton Mifflin	1978	1
137708	1845170423	Cocktail Classics	David Biggs	2004	Connaught	2004	1
137709	0449906736	Flashpoints: Promise and Peril in a New World	Robin Wright	1993	Ballantine Books	1993	1
137710	0440403968	There's a Bat in Bunk Five	Paula Danziger	1988	Random House Childrens Pub (Mm)	1988	1
137711	0525447644	From One to One Hundred	Teri Sloat	1991	Dutton Books	1991	1

137712 rows × 7 columns

4. Popular book of same_author

```
Enter_author_name:-A A Milne
```

	ISBN	Book-Title	Book-Author_x	Year-Of-Publication	Publisher	count
0	0525447083	Piglet Meets a Heffalump	A A Milne	2002	Penguin Putnam~childrens Hc	3
1	0525447148	Christopher Robin Gives Pooh a Party	A A Milne	2002	Penguin Putnam~childrens Hc	1
2	0525447105	Kanga and Baby Roo Come to the Forest	A A Milne	2002	Penguin Putnam~childrens Hc	1
3	0525447075	Pooh Goes Visiting and Pooh and Piglet Nearl	A A Milne	2002	Penguin Putnam~childrens Hc	1
4	0525459243	Winnie the Pooh Storybook Treasury	A A Milne	2002	Penguin Putnam~childrens Hc	1

5.Top most rating books.

```
Enter no of top rated books you want:- 5
```

Top 5 books

	ISBN	Book-Title	Book-Author	Year-Of-Publication	Publisher	no_of_user_review	Total_sum_ratings	Avarage_rating
307	0345339738	The Return of the King (The Lord of the Rings,...	J.R.R. TOLKIEN	1986	Del Rey	77	724	9.402597
348	0439139597	Harry Potter and the Goblet of Fire (Book 4)	J. K. Rowling	2000	Scholastic	137	1269	9.262774
94	0345339711	The Two Towers (The Lord of the Rings, Part 2)	J.R.R. TOLKIEN	1986	Del Rey	83	757	9.120482
380	0439136369	Harry Potter and the Prisoner of Azkaban (Book 3)	J. K. Rowling	2001	Scholastic	133	1208	9.082707
295	0064400557	Charlotte's Web (Trophy Newbery)	E. B. White	1974	HarperTrophy	68	617	9.073529

6.Popular Book of same Publsher

```
Enter_publisher_name:-Penguin Putnam-childrens Hc
```

	ISBN	Book-Title	Book-Author	Year-Of-Publication	Publisher_x	count
10	0670846503	Gerald's Game	Stephen King	1992	Penguin Putnam-childrens Hc	10
13	0140328718	James and the Giant Peach	Roald Dahl	1988	Penguin Putnam-childrens Hc	6
9	0895776839	Selections From Household Hints and Handy	Readers Digest	1994	Penguin Putnam-childrens Hc	5
2	0762102519	Kiss Before Dying (The Best Mysteries of All T...	Ira Levin	1999	Penguin Putnam-childrens Hc	5
1	0670886017	Kits Law	Donna Morrissey	1999	Penguin Putnam-childrens Hc	5
15	0670870129	Simplicity	Edward De Bono	2002	Penguin Putnam-childrens Hc	4
0	0670846549	Secret History	Donna Tartt	2002	Penguin Putnam-childrens Hc	4
5	0762101067	Best of Sisters In Crime	Marilyn Wallace	2002	Penguin Putnam-childrens Hc	3
8	0525447083	Piglet Meets a Heffalump	A A Milne	2002	Penguin Putnam-childrens Hc	3
7	0723215537	Observers Book of Aircraft 1976	William Green	1976	Penguin Putnam-childrens Hc	2
12	0670828076	Dream of Old Leaves	Bret Lott	1989	Penguin Putnam-childrens Hc	2
27	0895775255	Perry Mason Casebook	Erle Stanle Gardner	1993	Penguin Putnam-childrens Hc	2
32	0670826391	Art of Robert Bateman	Ramsay Derry	2002	Penguin Putnam-childrens Hc	2
3	0525447091	Eeyore Has a Birthday	A.a Milne	2002	Penguin Putnam-childrens Hc	2
35	0895770407	Fix It Yourself Manual	Readers Digest Staff	1977	Penguin Putnam-childrens Hc	2
33	0670857785	Idoru Uk Edition	William Gibson	2002	Penguin Putnam-childrens Hc	1
29	0670856975	Quarantine	Jim Crace	2002	Penguin Putnam-childrens Hc	1
30	0670000329	Death of a Salesman	Arthur Miller	2002	Penguin Putnam-childrens Hc	1
31	0670000094	Portrait of the Artist As a Young Man	James Joyce	1975	Penguin Putnam-childrens Hc	1
36	0525222650	The Tree Where Man Was Born : The African Expe...	Peter Matthiessen	1972	Penguin Putnam-childrens Hc	1
34	0895770105	Reader's Digest Complete Do It Yourself Manual	Readers Digest	1973	Penguin Putnam-childrens Hc	1
37	0525459243	Winnie the Pooh Storybook Treasury	A A Milne	2002	Penguin Putnam-childrens Hc	1
41	0670821705	Moved and the Shaken	Ken Dryden	2002	Penguin Putnam-childrens Hc	1
38	0762102624	Energize Your Life (Health and Healing the Nat...	Readers Digest	2001	Penguin Putnam-childrens Hc	1
39	0448165333	Second Twelve Months of Life	Frank Caplan	2002	Penguin Putnam-childrens Hc	1
40	0399523251	Party Fare: A Healthy Exchanges Cookbook	Joanna M. Lund	2002	Penguin Putnam-childrens Hc	1
28	0525249842	Sam Walton the Inside Story of Americas	Vance H Trimble	2002	Penguin Putnam-childrens Hc	1
21	0843103752	Ta for Tots and Other Prinzes	Alvyn M Freed	1976	Penguin Putnam-childrens Hc	1
26	0525447075	Pooh Goes Visiting and Pooh and Piglet Nearl	A A Milne	2002	Penguin Putnam-childrens Hc	1

7. Collaborative Filtering (User-Item Filtering)

```
Enter book name:-Harry Potter and the Goblet of Fire (Book 4)
Input Book:
Harry Potter and the Goblet of Fire (Book 4)

RECOMMENDATIONS:

Harry Potter and the Chamber of Secrets (Book 2)
Harry Potter and the Sorcerer's Stone (Book 1)
Harry Potter and the Prisoner of Azkaban (Book 3)
The Two Towers (The Lord of the Rings, Part 2)
The Fellowship of the Ring (The Lord of the Rings, Part 1)
```

8.Content Based filtering

```
Recommended Books:

Harry Potter and the Chamber of Secrets (Book 2)
Harry Potter and the Prisoner of Azkaban (Book 3)
Harry Potter and the Sorcerer's Stone (Book 1)
From Potter's Field
The Book of Ruth (Oprah's Book Club (Paperback))
```

5. Conclusion :

- In EDA, the Top-10 most rated books were essentially novels. Books like The Lovely Bone and The Secret Life of Bees were very well perceived.
- Majority of the readers were of the age bracket 30-40 and most of them came from North American and European countries namely USA, Canada, UK, Germany and Spain. Among them USA have highest no book.
- If we look at the ratings distribution, most of the books have high ratings with maximum books being rated 8. Ratings below 5 are few in number.
- In book Wild Animus user gives a high no of rating which is 2.5k its so high in compare to other books.

- Author with the most books was Agatha Christie, William Shakespeare and Stephen King.

- For modelling, it was observed that for model based collaborative filtering and content based filtering doing best .
- In other aproach the avg rating The Return of the King and hary poter have most liked books .