Capstone Project-4

Book Recommendation System

By- Shubham Srivastava

Index

- Defining Problem Statement
- Dataset Summary
- EDA(Exploratory Data Analysis)
- Imputing Missing values
- Evaluation using different Recommendation Models
- 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 become much more important in our lives in terms of providing highly personalized and relevant content.

The main objective is to create a recommendation system to recommend relevant books to users based on popularity and user interests.

Dataset Summary

The dataset is comprised of 3 csv files: User_df, Books_df, Ratings_df

User_dataset:

- User_ID(unique for each user)
- Location(contains city, state and country separated by commas)
- Age

Shape of dataset- (278858,3)

Books_dataset:

ISBN (unique for each book)

Book Title Book Author Year of Publication

Publisher Image-URL-S Image-URL-M

Image-URL-L

Shape of dataset- (271360,8)

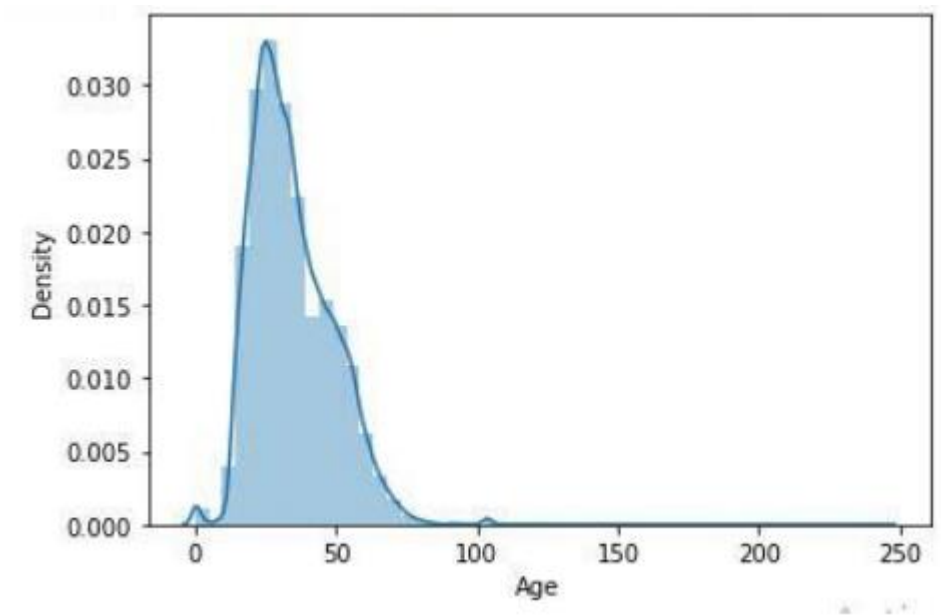
Ratings_dataset:

User_ID Book-Rating ISBN

Shape of dataset- (1149780,3)

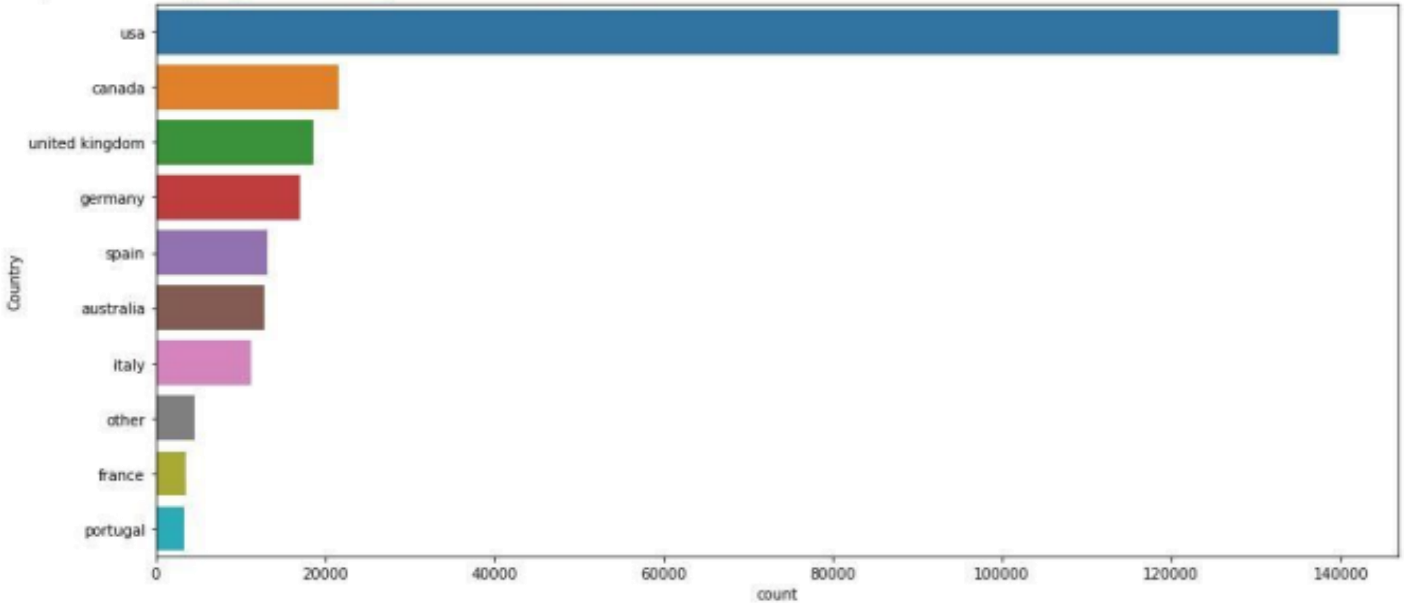
Observation from Users_df (Age)

- The Age range from 0 to 250 which is inconceivable.
- The Age range distribution is Right Skewed.
- Most active readers lie in Age group of 20-40.



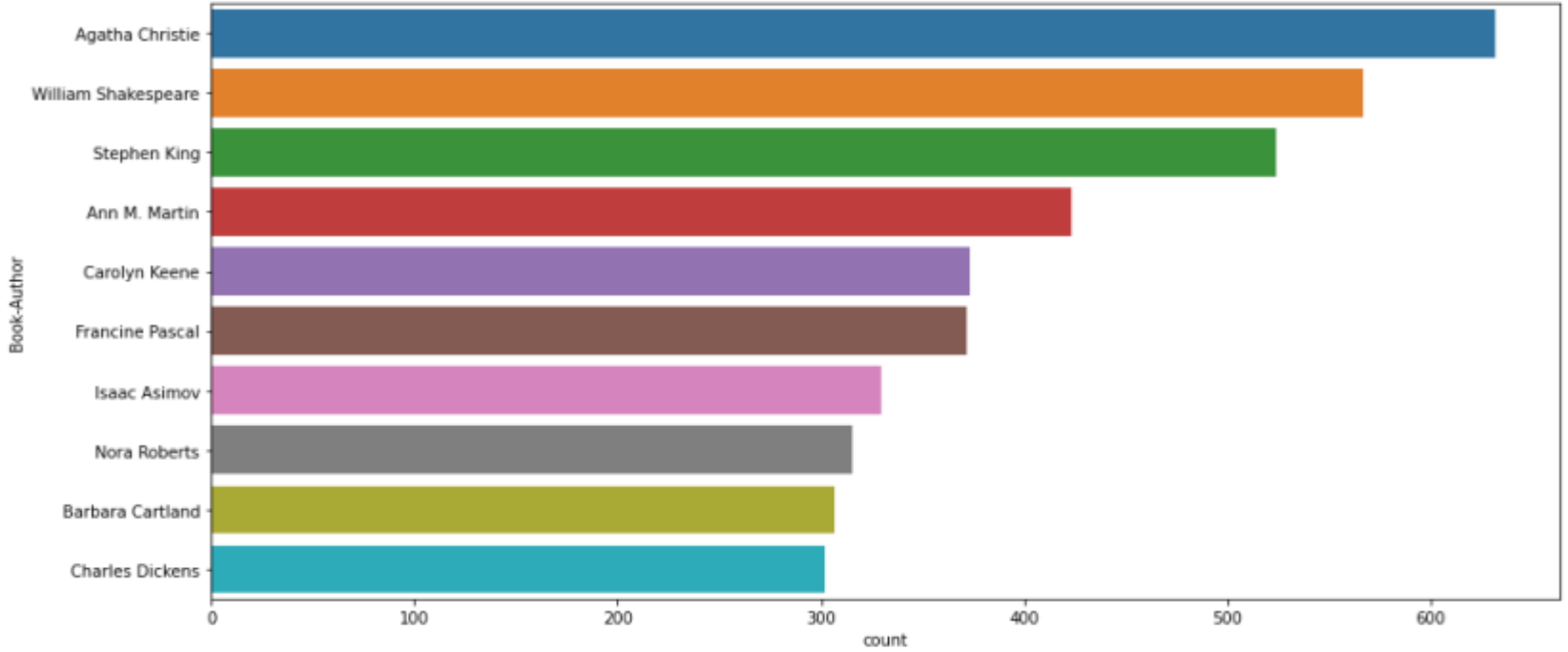
Observation from Users_df (Location)

- We split the location column and analyzed country.
- USA has most active readers.



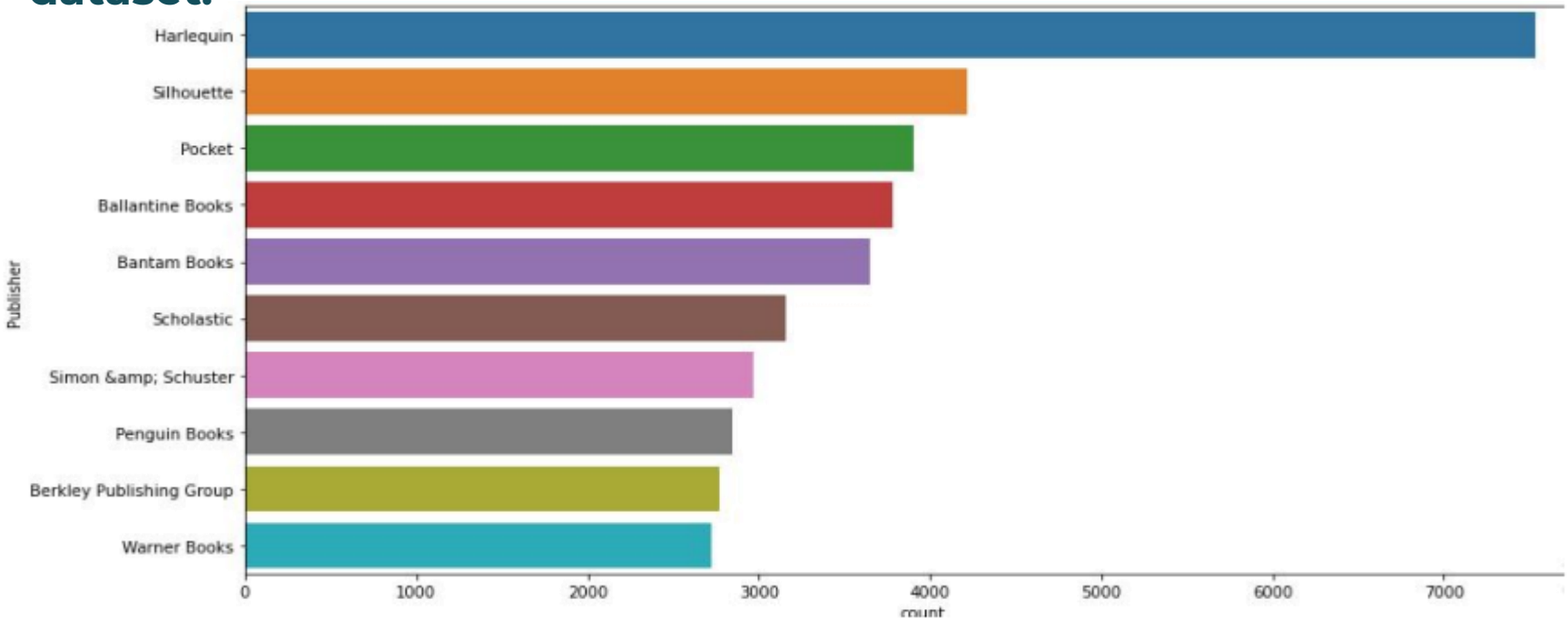
Observation from Book_df (Authors)

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



Observation from Book_df (Publishers)

- **Harle Quin published highest number of books in our given dataset.**

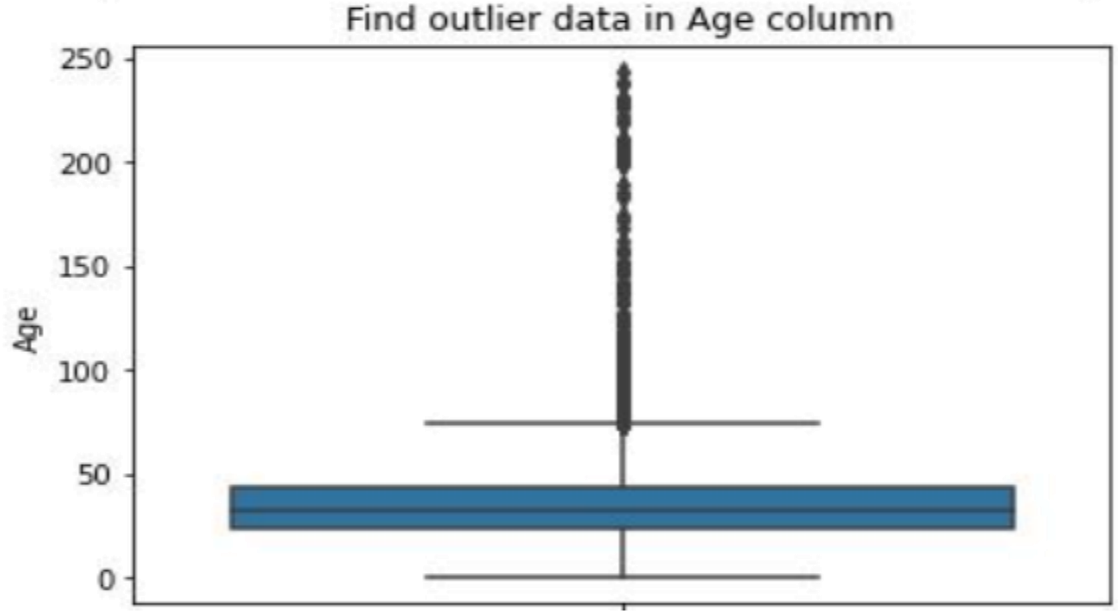


Observation from Ratings_df (Book_Rating)

- **Rating 8 has been rated the highest number of times.**

Data Cleaning

- Age column has 40% missing values.
- There are outliers in Age column.
- Age has positive skewness so we can use median to fill Nan values.



Different Models

- **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 chart
  R is the average rating of the book
  C is the mean vote across the whole report

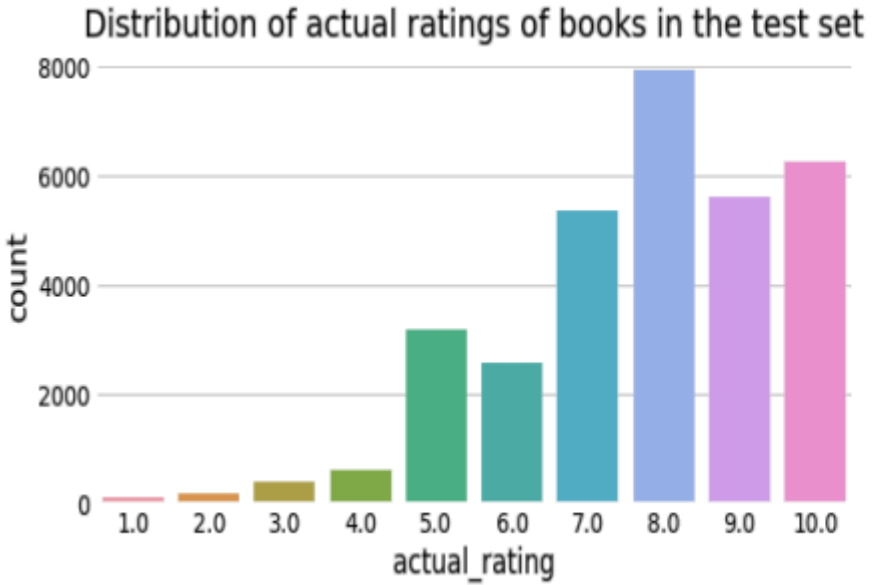
Different Models

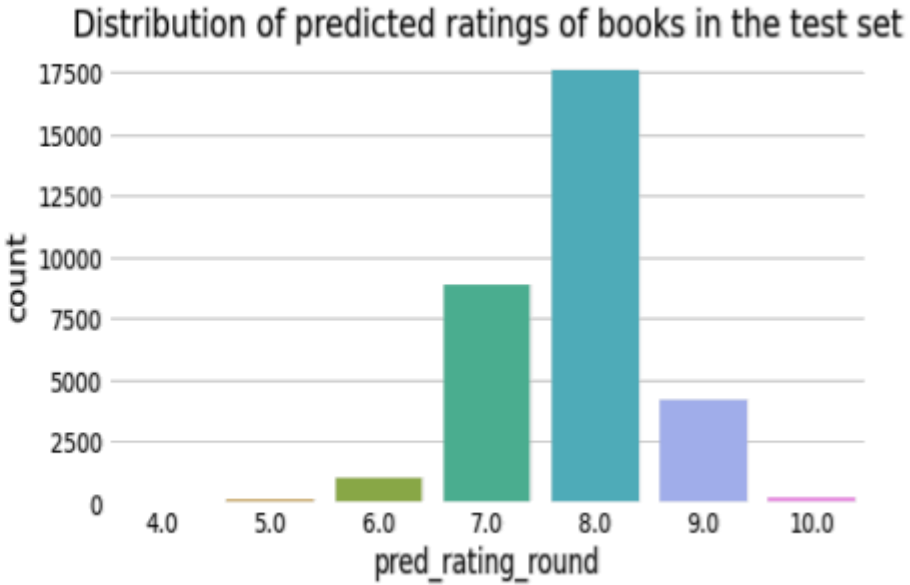
	Book-Title	Total_No_Of_Users_Rated	Avg_Rating	Score
0	Harry Potter and the Goblet of Fire (Book 4)	137	9.262774	8.741835
1	Harry Potter and the Sorcerer's Stone (Harry Potter (Paperback))	313	8.939297	8.716469
2	Harry Potter and the Order of the Phoenix (Book 5)	206	9.033981	8.700403
3	To Kill a Mockingbird	214	8.943925	8.640679
4	Harry Potter and the Prisoner of Azkaban (Book 3)	133	9.082707	8.609690
5	The Return of the King (The Lord of the Rings, Part 3)	77	9.402597	8.596517
6	Harry Potter and the Prisoner of Azkaban (Book 3)	141	9.035461	8.595653
7	Harry Potter and the Sorcerer's Stone (Book 1)	119	8.983193	8.508791
8	Harry Potter and the Chamber of Secrets (Book 2)	189	8.783069	8.490549
9	Harry Potter and the Chamber of Secrets (Book 2)	126	8.920635	8.484783
10	The Two Towers (The Lord of the Rings, Part 2)	83	9.120482	8.470128
11	Harry Potter and the Goblet of Fire (Book 4)	110	8.954545	8.466143
12	The Fellowship of the Ring (The Lord of the Rings, Part 1)	131	8.839695	8.441584
13	The Hobbit : The Enchanting Prelude to The Lord of the Rings	161	8.739130	8.422706
14	Ender's Game (Ender Wiggins Saga (Paperback))	117	8.837607	8.409441
15	Tuesdays with Morrie: An Old Man, a Young Man, and Life's Greatest Lesson	200	8.615000	8.375412
16	Charlotte's Web (Trophy Newbery)	68	9.073529	8.372037
17	Dune (Remembering Tomorrow)	75	8.973333	8.353301
18	A Prayer for Owen Meany	181	8.607735	8.351465
19	Fahrenheit 451	164	8.628049	8.346969

Different Models

- **Model Based Collaborative Filtering**

SVD model results





Different Models

- **Collaborative Filtering (Item –Item Based)**
  Cosine Similarity
  Nearest Neighbor

```
Recommendations for Angels &amp;amp; Demons:

1: The Da Vinci Code, with distance of 0.8275555141289059:
2: Digital Fortress : A Thriller, with distance of 0.83781217691282:
3: Deception Point, with distance of 0.8422605379839627:
4: Prey: A Novel, with distance of 0.9216969275206289:
5: The Cat Who Knew a Cardinal, with distance of 0.9280814355076102:
```

Different Models

- **Collaborative Filtering (User –Item Based)**

```
Enter User ID from above list for book recommendation  69078
Recommendation for User-ID =  69078
         ISBN                                          Book-Title  recStrength
0  0446310786                                To Kill a Mockingbird        0.842
1  0345370775                                        Jurassic Park        0.802
2  0312966970          Four To Score (A Stephanie Plum Novel)         0.675
3  0316769487                              The Catcher in the Rye         0.673
4  0345361792                           A Prayer for Owen Meany           0.646
5  0440214041                                 The Pelican Brief           0.621
6  044021145X                                         The Firm           0.617
7  0440211727                                      A Time to Kill         0.617
8  0060928336  Divine Secrets of the Ya-Ya Sisterhood: A Novel          0.606
9  0312924585                                Silence of the Lambs         0.600
```

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 20-35 and most of them came from North American and European countries namely USA, Canada, UK, Germany and Spain.
- 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.
- Author with the most books was Agatha Christie, William Shakespeare and Stephen King.
- For modelling, it was observed that for model based collaborative filtering SVD technique worked way better than NMF with lower Mean Absolute Error (MAE)

Conclusion

A recommendation system helps an organization to create loyal customers.
The recommendation system today are very 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

Challenges

- Handling of sparsity was a major challenge as well 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 Location etc..

- Decision making on missing value imputations and outlier treatment was quite challenging as 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