UnSupervised ML Project

Book Recommender System

By-Aditi Rajguru

Why do we need a book Recommender?

Book reviews give books greater visibility and a greater chance of getting found by more readers.

On some websites, books that have more book reviews are more likely to be shown to prospective readers and buyers as compared to books with few or no book reviews.

Book reviews also help amplify your book’s reach among book clubs, bookstores, blogging communities and other opportunities to gain attention from new readers.

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.

Contents

- Datasets used and Feature Representation
- Introduction
- Feature Engineering
- Feature Transformation
- Exploratory Data Analysis with Data Visualization
- Building Recommender
- Challenges Faced
- Conclusion

Datasets used

➔ Books_df:This dataset contains all the information about book like its id,title,year of publication,publisher,etc.

➔ Users_df: This dataset contains demographic information of users by theirs id's , location and age.

➔ Ratings_df:This dataset contains ratings of books which is either explicit 1-10 or implicit which is 0.

Feature Representation:

- ➢ Books_df:
- ISBN: For identifying book its unique ISBN number.
- Book-Title:Title of the book.
- Book-Author:Author of the book.
- Year-Of-Publication:The year in which each book was published.
- Publisher:Publisher of the book.
- The above features are obtained from amazon web services and in case of several authors only first is provided.While the below given image urls are linkings to cover pages appearing three different flavors,pointing to Amazon website.
- Image-URL-S:Image url in small.
- Image-URL-M:Image url in medium
- Image-URL-L:Image url in Large.

➢ Users_df:

- User-ID:Anonymized unique id’s for users and mapped to integers.
- Location:location if provided by user.
- Age:Age of the user.

Location and age of the users are only provided if available otherwise these fields contains null values.

➢ Ratings_df:

- Ratings:Contains book rating information which is either explicit expressed on scale of 1-10 or implicit expresses by zero.

Feature Engineering:

- **Duplicate Values:**
- There were no duplicate values in our dataset.

- **Missing Values:**
- For this Unsupervised project there were minor amount missing values in Books_df which was later removed and there were huge amount of missing values in Users_df age column.which were later well sorted and filled with valid values.

Feature Transformation:

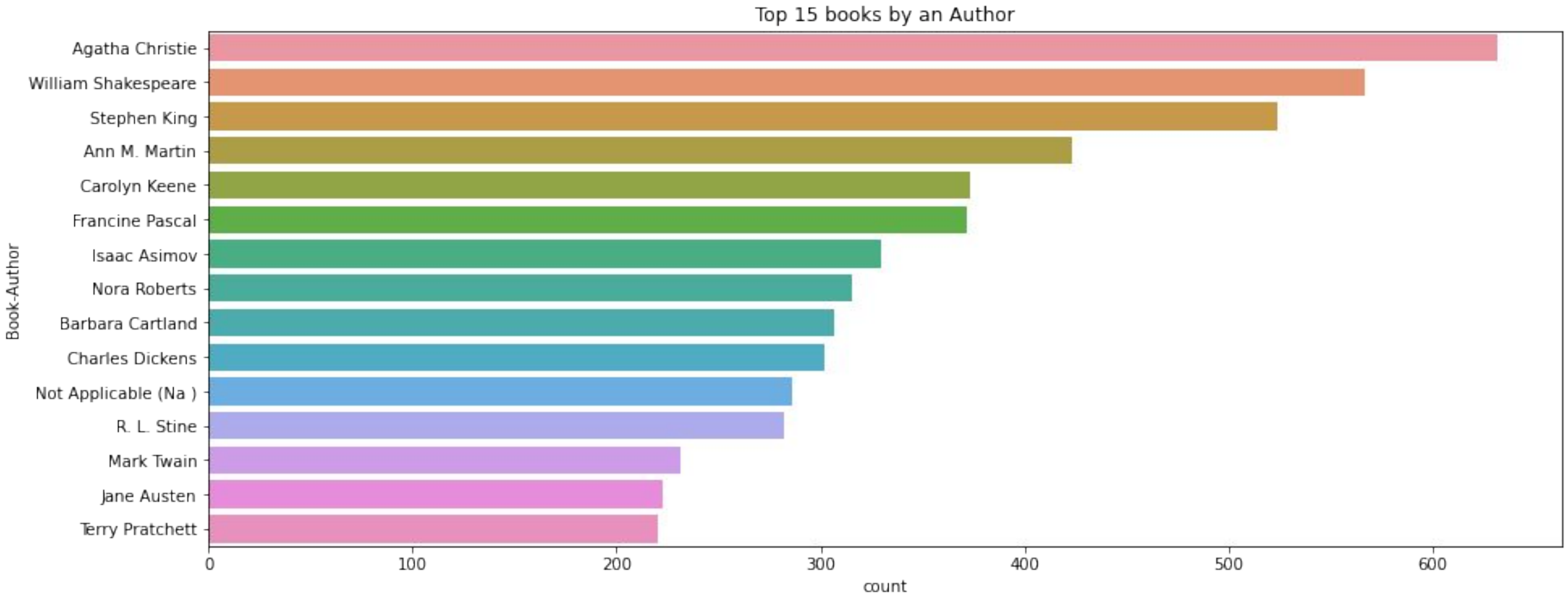
- ★ Age column was fully transformed into valid ages by eliminating all the age below 8 and above 90.
- ★ In Book_df all the image urls were removed since they were of no use for this project and for further analysis and other columns were renamed.
- ★ As well as In User_df and Ratings_df also columns were rename so that we can access them easily for further operations.

Exploratory Data Analysis:

Why EDA is important?

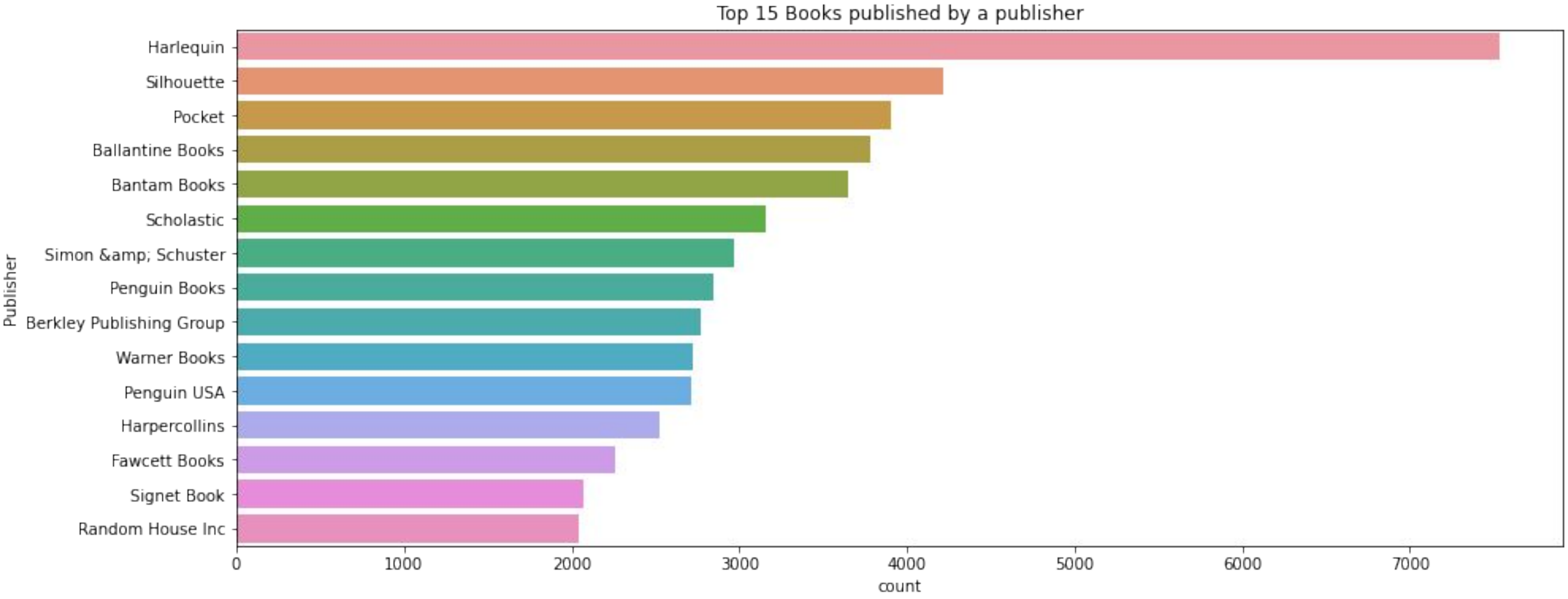
For an understanding of your data, it’s characteristics, and it’s distributions is vital to any successful data science task, whether it’s inference or prediction. And contrary to what you might expect, the reason for EDA’s importance is not technical, and has nothing to do with programming. It’s the thing that separates a mediocre data scientist from a great one — decisions.

Top 15 books by an Author:



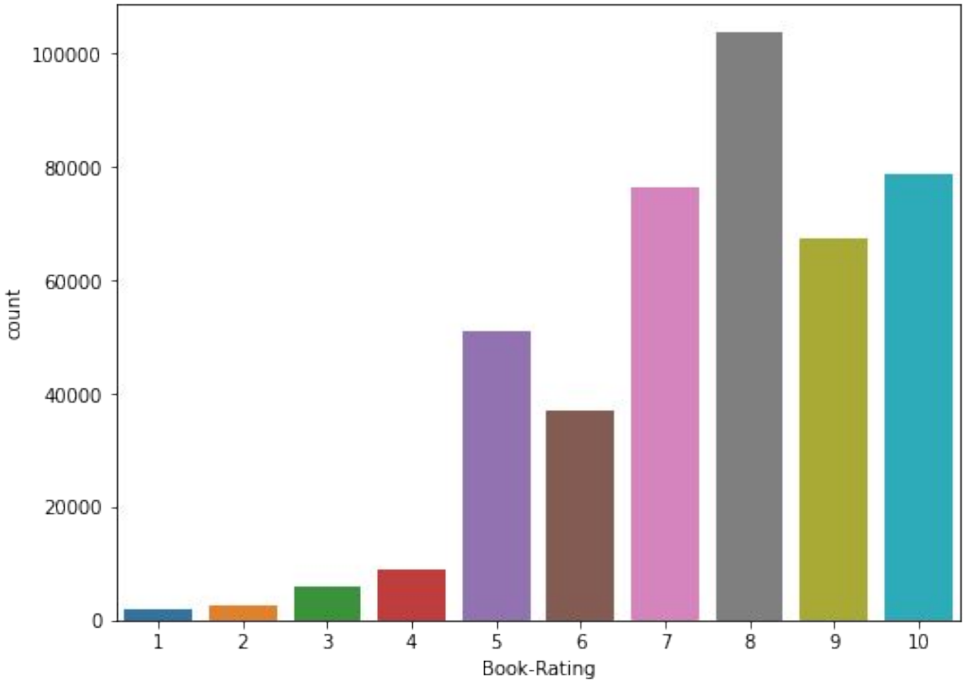
- In top 15 authors Agatha Christie tops the list while Mark Twain,Jane Austen and Terry Prachett are in the bottom of the list

Top 15 Books published by a publishers:



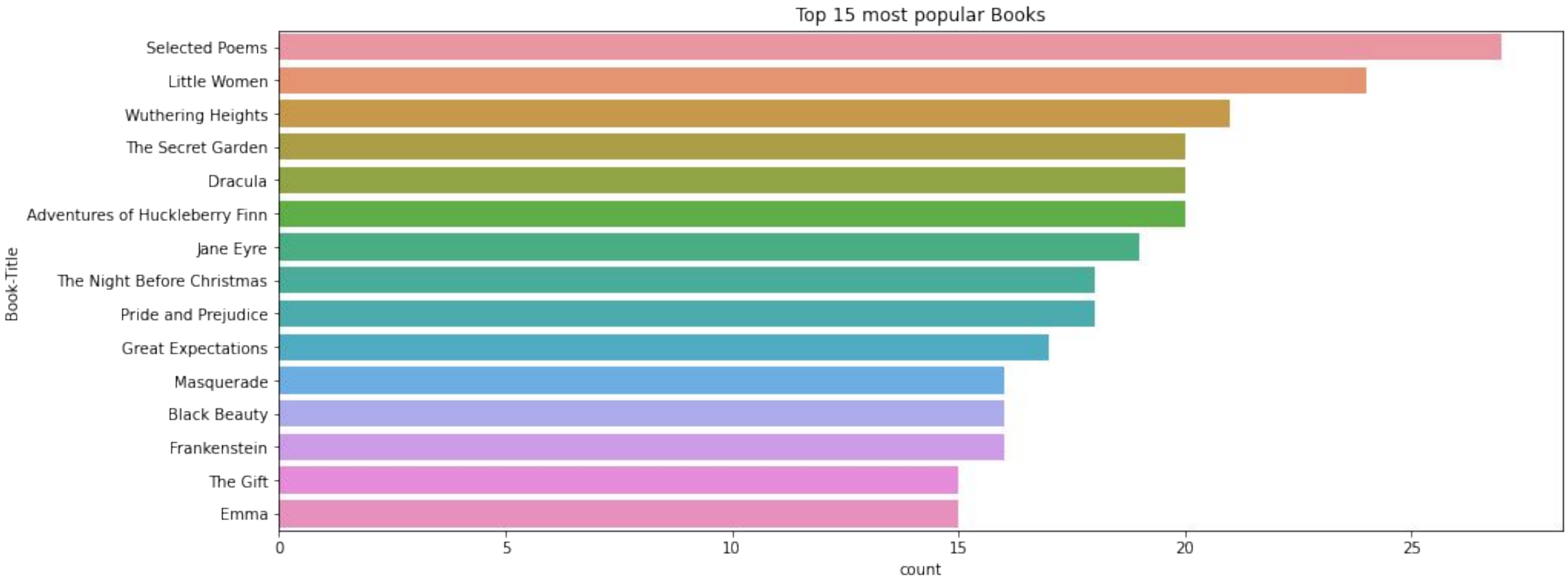
Top 15 books published by a publisher has Harlequin at top and Random House Inc at the bottom.

Distribution of ratings on books:



Most books are rated above 7 and as we can see in figure all books has mixed-kind of ratings.

Top 15 most popular Books:



Selected Poems are in the top of 15 most popular books.While The gift,Emma and Frankenstein are in bottom five books.

Recommendation System:

Unsupervised techniques used for this project:

Collaborative Filtering:

Collaborative filtering is a technique that can filter out items that a user might like on the basis of 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.

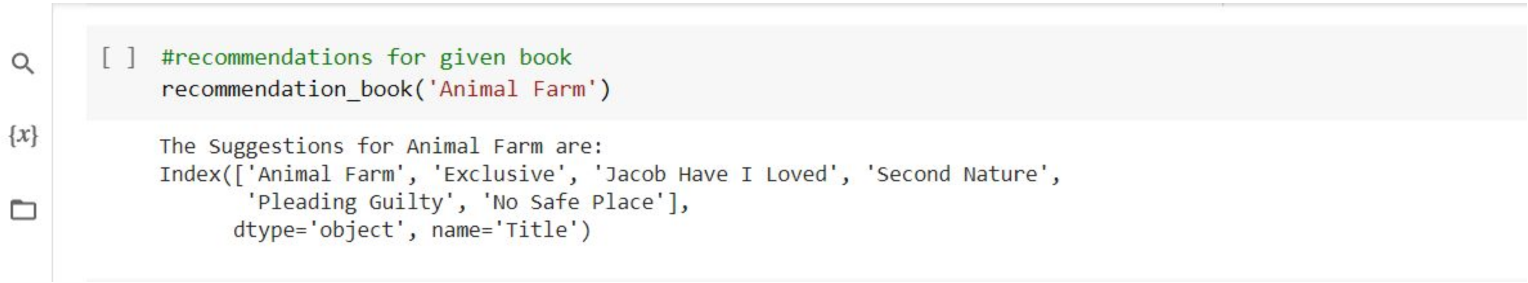
K-Means Algorithm:

K-means clustering is the unsupervised machine learning algorithm that is part of a much deep pool of data techniques and operations in the realm of Data Science. It is the fastest and most efficient algorithm to categorize data points into groups even when very little information is available about data

Approach:

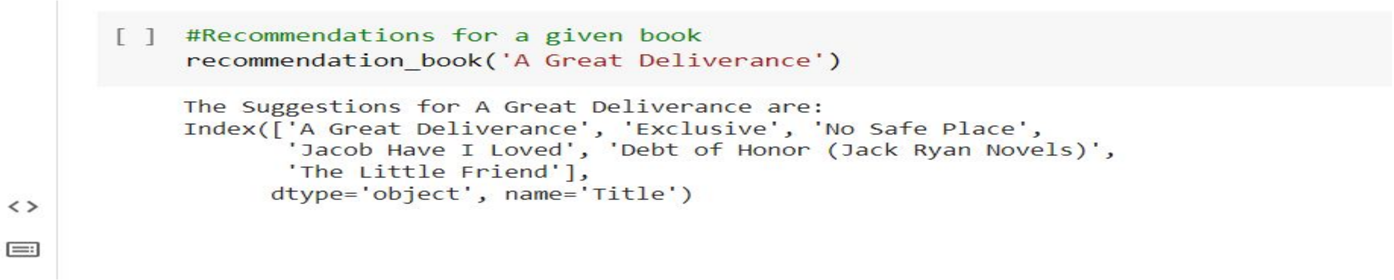
Here we have used only those Users which have rated more than 200 books and books which has ratings more than 50.The main motive to do so is to ensure that recommender engine should give appropriate results.Because all readers don’t review books and sometimes review given by person don’t always have considerable reviews.So we decided to go with above mentioned data to find those users which are actual readers and can give appropriate ratings and only those books which has more than 50 rating so that in future it will be of some use.

Recommendations:For Books Animal Farm and A Great Deliverance



```
[ ]  #recommendations for given book
     recommendation_book('Animal Farm')

     The Suggestions for Animal Farm are:
     Index(['Animal Farm', 'Exclusive', 'Jacob Have I Loved', 'Second Nature',
            'Pleading Guilty', 'No Safe Place'],
           dtype='object', name='Title')
```



```
[ ]  #Recommendations for a given book
     recommendation_book('A Great Deliverance')

     The Suggestions for A Great Deliverance are:
     Index(['A Great Deliverance', 'Exclusive', 'No Safe Place',
            'Jacob Have I Loved', 'Debt of Honor (Jack Ryan Novels)',
            'The Little Friend'],
           dtype='object', name='Title')
```

Conclusion:

- Most of the people knows William Shakespeare but in top 15 books by author Agatha Christie ranks above William Shakespeare.

- JK Rowling most popular Author among teens does not make it to top 15 list. And also there are more books which are popular than harry potter over the years it keeps on changing.

- Very few books have ratings 10 while most of the books are not even rated.

- Little women ,An American novel ranks on second place in top 15 most popular Books.

- Some Books were having more than more volumes and different authors.

- Pride and Prejudice ranks in top 10 most popular books which also have movie based on it.

Challenges faced:

- **Lack of Data:** Perhaps the biggest issue facing recommender systems is that they need a lot of data to effectively make recommendations.A good recommender system firstly needs item data (from a catalog or other form), then it must capture and analyze user data (behavioral events), and then the magic algorithm does its work. The more item and user data a recommender system has to work with, the stronger the chances of getting good recommendations. But it can be a chicken and egg problem – to get good recommendations, you need a lot of users, so you can get a lot of data for the recommendations.

- Huge data:The given data is very huge and as huge the data is more is its complexity,We faced same for our project.Data was very huge and it also had lot's of missing values and also some information which was irrelevant. In case of this project.

- Computation time required to work on datasets was more since it was a bit tough dataset.

- Missing values were having mixed kind of impact on project with different datasets.So deciding which approach was suitable took time in trying different approaches