Book recommendation system

Nakshith D N
Data science trainee,
AlmaBetter, Bangalore

Abstract

The Book Recommendation System aims to provide the best suggestion to the user by analysing the buyer's interest. The quality and the content are taken into consideration by employing content filtering, association rule mining and collaborative filtering. The booming technology of the modern world has given rise to the enormous book websites. This makes the buyers choose the best books to read as books play a vital role in many people's lives. Various kinds of books come into existence on a day to day basis. So in order to eliminate this critical situation the recommendation system has been introduced in which the suggestion on the various books can be provided based on the analysis of the buyer's interest. The Book Recommendation System is an intelligent algorithm which reduces the overhead of the people. This provides benefits to both the seller and the consumer creating the win-win situation. The E-commerce site to network security, all demands the need for the recommended system to increase their revenue rate. The content filtering, association rule mining and collaborative filtering are the various decision making techniques employed in the recommendation system as it helps buyers by the strong recommendations as there are various books, buyer's sometimes cannot find the item they search for. The Book Recommendation System is widely implemented using search engines consisting of data sets. The collaborative filtering involves the analysis of the opinions in which the recommendation is provided based on the ratings provided by the users. The quality of the item cannot be analysed in the content based filtering. But the collaborative filtering can expose the quality of the item. The collaborative filtering is employed in two ways namely, the user based collaborative filtering and item based collaborative filtering.

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.

Introduction

Recommender systems have become a part of daily life for users of Amazon and Netflix and even social media. While some sites might use these systems to improve the customer experience (if you liked movie A, you might like movie B) or increase sales (customers who bought product C also bought product D), others are focused on customised advertising and suggestive marketing. As a book lover and former bookstore manager, I have always wondered where I can find good book recommendations that are both personalised to my interests and also capable of introducing me to new authors and genres. The purpose of this project is to create just such a recommender system (RS).

Dataset

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 flavours (Image-URL-S,

Image-URL-M, Image-URL-L), i.e., small, medium, large. These URLs point to the Amazon website.

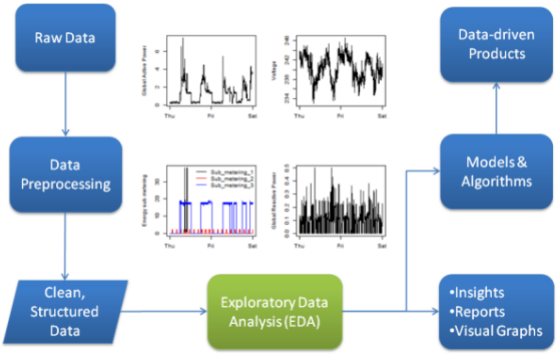
• Ratings

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,

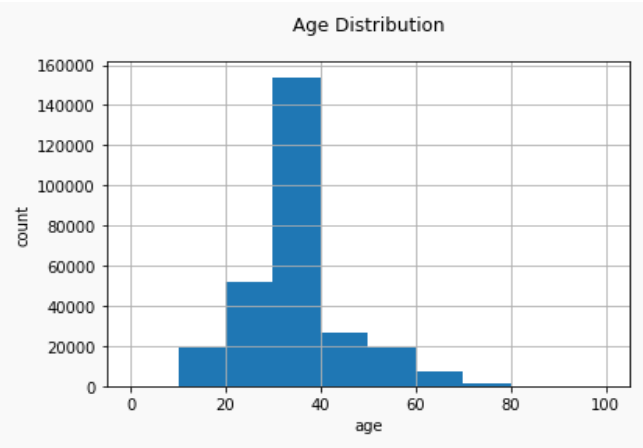
expressed by 0.

Architecture

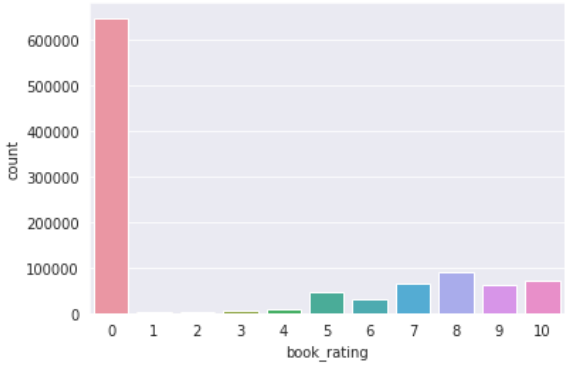


Data Exploration and visualisation

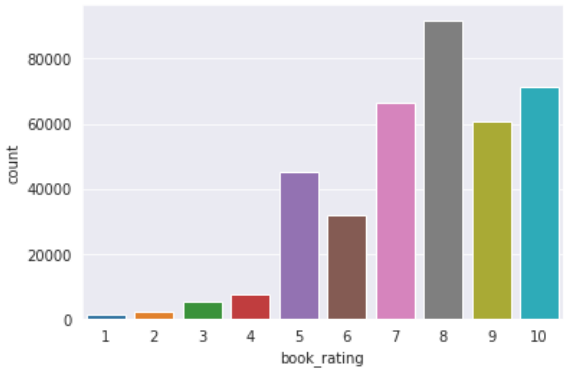
The purpose of exploratory data analysis is to identify the variables that impact payment default likelihood and the correlations between them. We use graphical and statistical data exploratory analysis tools to check every categorical variable. Each starts with a visualisation and is followed by a statistical test to verify the findings.



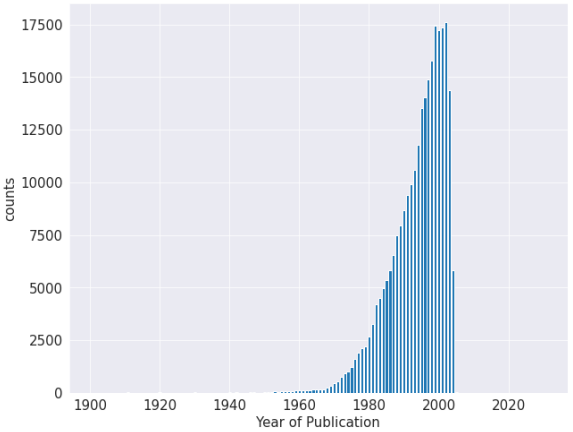
Here we can clearly observe the age distribution of the readers and we can clearly see that the age group of around 30-40 years are higher compared to all other age groups.



This countplot shows users have rated 0 the most, which can mean they haven't rated books at all. We have to separate the explicit ratings represented by 1–10 and implicit ratings represented by 0.



Now this countplot of bookRating indicates that higher ratings are more common amongst users and rating 8 has been rated highest number of times.



Here we can see publication years are somewhat between 1950 - 2005 here and there is a drastic increase in the publication from the beginning of the year 2000.

Collaborative Filtering using K-Means

k-means clustering is a method of vector quantization, originally from signal processing, that aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean (cluster centres or cluster centroid), serving as a prototype of the cluster. This results in a partitioning of the data space into Voronoi cells. k-means clustering minimises within-cluster variances (squared Euclidean distances), but not regular Euclidean distances, which would be the more difficult Weber problem: the mean optimises squared errors, whereas only the geometric median minimises Euclidean distances.

We want to get the information whether or not a person likes that book for our filtering. Criteria for that will be : If a person rates a book more than his/her average rating then he/she likes the book.

Creating crosstab for each user and each book and using the pivot table we have created a dataframe for the easier analysis and interpretation

Later in order to reduce the dimensions we used principal component analysis.

Since we are using k-means we need find the value of k, elbow method and silhouette analysis are used to find the value of k, since silhouette analysis has better results compared to elbow method we have used silhouette analysis and after applying the following results are obtained

For n_clusters = 2 The average silhouette_score is : 0.7246440630264265

For n_clusters = 3 The average silhouette_score is : 0.6711916007114602

For n_clusters = 4 The average silhouette_score is : 0.6682637986000894

For n_clusters = 5 The average silhouette_score is : 0.5885817459186526

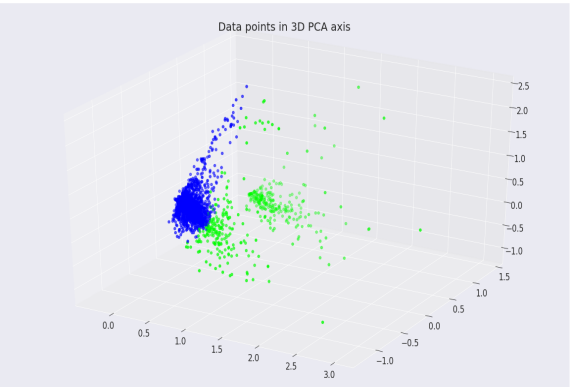
For n_clusters = 6 The average silhouette_score is : 0.6037061232528691

For n_clusters = 7 The average silhouette_score is : 0.5966723111929981

For n_clusters = 8 The average silhouette_score is : 0.5264456810087846

From the above results we can clearly say that the value which is highest provides good results so for n_clusters = 2 the optimal results can be obtained, it seems k=2 provides the best clustering.

Applying Clustering



After applying clustering we can clearly see that the data points are separated as shown in the previous 3d graph,

So now we need to analyse the each clusters for knowing the results

From the first cluster

The following results from the analysis are obtained ,the median Year was 1997,

The top 5 Books are,

1. The Da Vinci Code

2. Harry Potter and the Sorcerer's Stone (Harry Potter (Paperback))

3. Interview with the Vampire

4. Angels & Demons

5. Harry Potter and the Order of the Phoenix (Book 5)

And the top 5 Authors are,

1. Dan Brown

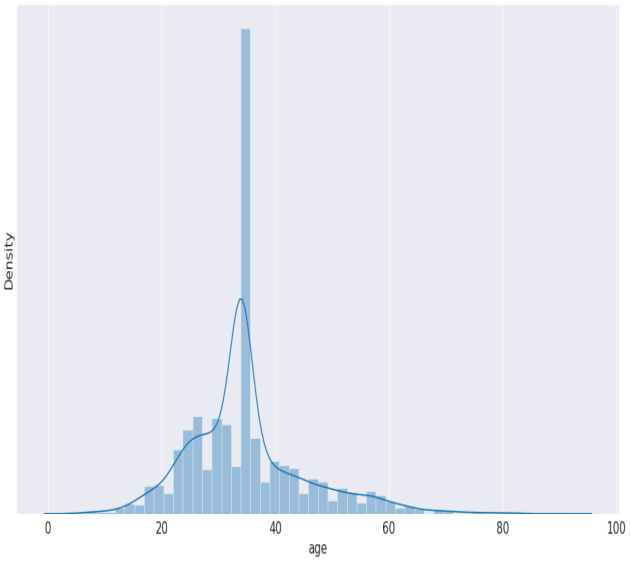
2. J. K. Rowling

3. Anne Rice

4. Anita Diamant

5. Michael Crichton

Most Common Location: toronto, ontario, canada, and the mean Age of the users was around 34.848339335734295.



From the second cluster

The following results are obtained from the second cluster and the Median Year was 1997

The top 5 Books are,

1. The Lovely Bones: A Novel

2. Where the Heart Is (Oprah's Book Club (Paperback))

3. The Secret Life of Bees

4. The Red Tent (Bestselling Backlist)

5. The Da Vinci Code

The top 5 Authors are,

1. Alice Sebold

2. Billie Letts

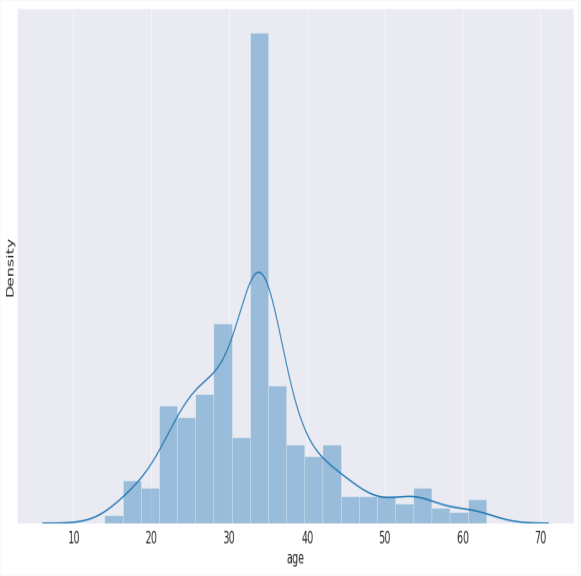
3. Sue Monk Kidd

4. Anita Diamant

5. Dan Brown

Most Common Location: chicago, illinois, usa

The mean Age of users is 33.617777777777775



Conclusion

- ★ The main objective which was to create a book recommendation system for users has been successfully done.
- ★ With collaborative filtering using K-Means we have successfully obtained the results.
- ★ The top 5 books and authors for each type of audience have been obtained successfully for specific user types.
- ★ By using two clusters we have obtained the results and they are clearly explained using respective visualisations.