

Capstone Project

Book Recommendation System

Rohit Madame

Abstract:

A recommendation system is one of the top applications of data science. Every consumer Internet company requires a recommendation system like Netflix, YouTube, a news feed, etc. What you want to show out of a huge range of items is a recommendation system. A recommendation engine is a class of machine learning which offers relevant suggestions to the customer. Before the recommendation system, the major tendency to buy was to take a suggestion from friends. But Now Google knows what news you will read, YouTube knows what type of videos you will watch based on your search history, watch history, or purchase history.

Problem Statement:

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.

The main objective is to predict the same genre books.

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

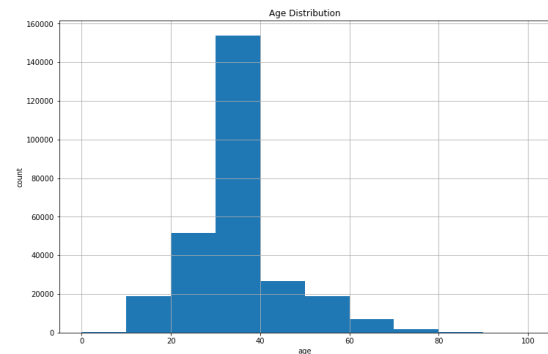
Steps Involved:

- ❖ Data Exploration and visualisation
- ❖ Model used to build recommendation system
- ❖ Challenges
- ❖ Conclusion

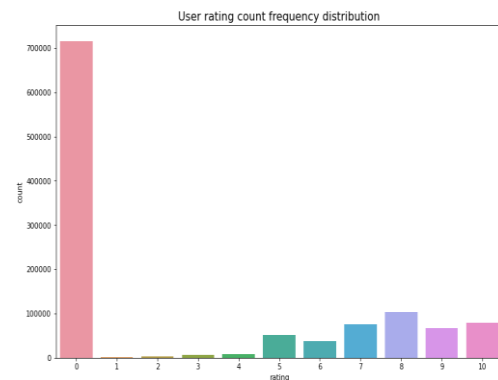
Data Exploration and visualisation:

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.

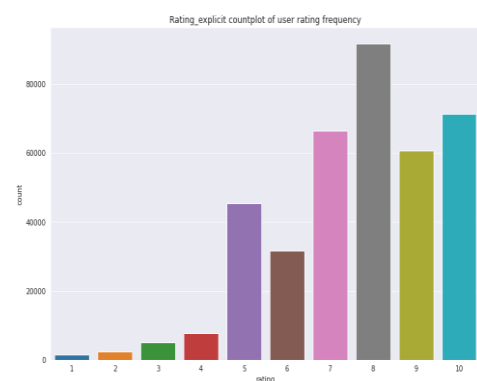
Here we can clearly observe the age distribution of the readers and we can clearly see that the age group of around 30-40 years are higher compared to all other age groups.



This count plot 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 book Rating indicates that higher ratings are more common amongst users and rating 8 has been rated number of times.



Model used to build recommendation system:

- ❖ **Popularity Based Recommendation(Base model)**
- ❖ **Collaborative Filtering**

Popularity Based Recommendation(Base model)

- ❖ It is a type of recommendation system which works on the principle of popularity and or anything which is in trend. These systems check about the books which are in trend or are most popular among the users and directly recommend them.
- ❖ I have sorted top 10 books from the dataset on the basis of number of ratings Received.

Collaborative Filtering

- ❖ Collaborative Filtering is considered to be one of the very smart recommender systems that work on the similarity between different users and also items that are widely used as an e-commerce website and also online movie websites. It checks about the taste of similar users and makes recommendations.
- ❖ The similarity is not restricted to the taste of the user, moreover there can be consideration of similarity between different items also. The system will give more efficient recommendations if we have a large volume of information about users and items.

1. Using KNN

- ❖ For this model we used KNN to find clusters of similar users based on common book ratings, and make predictions using the average rating of top-k nearest neighbors.

- ❖ In this model we assume that users who have given ratings more than 200 are users who read at least 20 books(suppose on user given rating 10/10 so minimum he read books($200 \text{ ratings} / 10 \text{ ratings per book} = 20$).For statistical significance we should consider only the data of user who given more than 200 ratings.
- ❖ To ensure statistical significance, users with less than 200 ratings, and books with less than 100 ratings are excluded.

2. Using SVD

- ❖ The SVD(Singular Value Decomposition) is used as a collaborative filtering technique. It uses a matrix structure where each row represents a user, and each column represents an item. The elements of this matrix are the ratings that are given to items by users.

Challenges:

- ❖ Since the data consisted of text data, data cleaning was a major challenge in feature engineering.
- ❖ Understanding the metric for evaluation was a challenge as well.

Conclusion:

- ❖ Recommendation system is unturned to exist in the e-commerce businesses with the help of collaborative or content-based filtering to predict different items and yes, users are most satisfied with the products recommended to them.
- ❖ As we have seen the age group between 30 to 40 reads the most books.
- ❖ We have removed the readers who did not give rating and then we found that most of the users gave 8 stars.

❖ KNN model gives good recommendation for books.

References:

1: Almbetter

2: GeeksforGeek

3: Analytics Vidhya