**Unsupervised ML Classification**

**Capstone Project**

**Book Recommendation System**

**Contributor Role:**

**Suraj Kad**

**(**[**Suraj.kad.90@gmail.com**](mailto:Suraj.kad.90@gmail.com)**)**

### **1. Data Cleaning and Preprocessing**

### **Libraries Used:**

* ipython-notebook - Python Text Editor
* sklearn - Machine learning library
* seaborn, matplotlib - Visualization libraries
* numpy, scipy- number python library
* pandas - data handling library

The dataset consists of three tables; Books, Users, and Ratings. Data from all three tables are cleaned and preprocessed separately as defined below briefly:

**For Books Table:**

* Drop Two Image URL features.
* Check for the number of null values in each column. There are only 3 null values in the table. Replace these three empty cells with ‘Other’.
* Check for the unique years of publications. Two values in the year column are publishers. Also, for three tuples the name of the author of the book was merged with the title of the book. Manually set the values for these three above obtained tuples for each of their features using the ISBN of the book.
* Upper-casing all the alphabets present in the ISBN column and removal of duplicate rows from the table.

**For Users Table:**

* Check for null values in the table. The Age column has more than 1 lakh null values.
* Check for unique values present in the Age column. There are many invalid ages present like 0 or 244.
* By keeping the valid age range of readers as 10 to 100, replace null values and invalid ages in the Age column with the mean of valid ages.
* The location column has 3 values: city, state, and country. These are split into 3 different columns named; City, State, and Country respectively. In the case of null value, ‘other’ has been assigned as the entity value.
* Removal of duplicate entries from the table.

**For Ratings Table:**

* Check for null values in the table.
* Check for the Rating column and User-ID column to be an integer.
* Removal of punctuation from ISBN column values and if that resulting ISBN is available in the book dataset only then consider dropping that entity.
* Upper-casing all the alphabets present in the ISBN column.
* Removal of duplicate entries from the table.

### **2. Algorithms Implemented:**

#### **Popularity Based Recommendation :**

* Actual Popularity Based top rated books
* Author Based Recommendation Systems
* Books By the Same Author, Publisher of Given Book Name

#### **User-Item Collaborative Filtering Recommendation**

Collaborative Filtering Recommendation System works by considering user ratings and finds cosine similarities in ratings by several users to recommend books. To implement this, we took only those books' data that have at least 50 ratings in all.

#### **Correlation Based Recommendation**

For this model, we have created the correlation matrix considering only those books which have total ratings of more than 50. Then a user-book rating matrix is created. For the input book using the correlation matrix, top books are recommended.

#### **Nearest Neighbor Based Recommendation**

To train the Nearest Neighbours model, we have created a compressed sparse row matrix taking ratings of each Book by each User individually. This matrix is used to train the Nearest Neighbours model and then to find n nearest neighbors using the cosine similarity metric.

## **Conclusion:**

In EDA, the Top-10 most rated books were essentially novels. Books like The Wild Animus and The Lovely Bones: A Novel

.

Majority of the readers were of the age bracket 20-50 and most of them came from North American and European countries namely USA, Canada, UK.

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 Stephen King, Nora Roberts and James Patterson.

A recommendation system helps an organization to create loyal customers.

The recommendation systems today are so 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.

**GitHub Link:**

**https://github.com/SurajKad/Mobile-Price-Range-Prediction.git**

**Drive https://drive.google.com/drive/folders/1PoKC7Q7NXAFena7F0NeSd6zhvexjknr7?usp=share\_link**