Book Recommendation System

Shivani Thakur

Data science trainees
Alma Better, Bangalore

ABSTRACT:

Today the amount of information in the internet growth very rapidly and people need some instruments to find and access appropriate information. One of such tools is called recommendation system. Recommendation systems help to navigate quickly and receive necessary information. Generally they are used in Internet shops to increase the profit. This paper proposes a quick and intuitive book recommendation system that helps readers to find appropriate book to read next. The overall architecture is presented with it's detailed description. We used a collaborative filtering method based on Pearson correlation coefficient. Finally the experimental results based on the online survey are provided with some discussions

Keywords: *predictive, Logistic regression, random forest, decision trees, XG boost, KNN, overfit, best model*

INTRODUCTION:

Recommendation systems were evolved as intelligent algorithms, which can generate results in the form of recommendations to users. They reduce the overhead associated with making best choices among the plenty. Now, Recommender systems can be implemented in any domain from E-commerce to network security in the form of personalized services. They provide benefit to both the consumer and the manufacturer, by suggesting items to consumers, which can't be demanded until the recommendations. Every recommender system comprises of two entities, one is user and other is item. A user can be any customer or consumer of any product or items, who get the suggestions. Input to recommendation algorithm can be a database of user and items and output obliviously will be the recommendations. Input for this system is customers and book data and output of this book denotes the book recommendations. Recommendation systems are used in hundr eds of different services everywhere from on line shopping to music to movies. Recomme nder systems are really critical in some indus tries as they can generate a huge amount of i ncome when they are efficient or also be a w ay to stand out significantly from competitos .

PROBLEM STATEMENT:

Develop a Book recommendation system using different recommendations models:

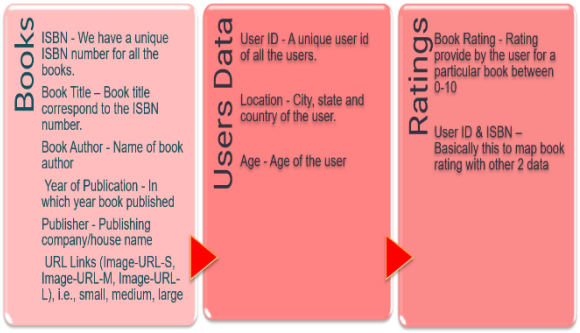
The main objective of our project is to create book recommendation systems for users on various approaches.

.

DATA DESCRIPTION:

The data contains information regarding Books ,Rating and Users. In different data set. The various features and information can be used to recommend the best book to the user. The data description phase starts with a initial data collection and proceeds with activities in order to get familiar with thedata. Identifying data quality problems,discovering first insights into the data anddetecting interesting subsets form hypotheses from hidden information areactivities of this step. The data was taken from online website .

We have 3 files in our dataset which is extracted from some books selling websites:



DATA PRESPROCESSING

Statistical analysis and descriptive statistics are two of the most crucial steps in the process. Developing predictive models, evaluating the models, and calculating their accuracy are the main tasks involved in this process. Irregularities are of different types of data.

- Missing Values
- Duplicate Values
- Anomalies/Outliers

DATA TRANSFORMATION

Data transformation is the process of normalizing and aggregating the data to further improve the efficiency and accuracy of data mining.

DATA PREPROCESSING:

Dataset may contain noise, missing values and inconsistent data. Thus, pre-processing of data is essential to improve the quality of data and time required in the data mining.

HANDLING OUTLIERS:

Outliers are data points that diverge from other observations for several reasons. During the EDA phase, one of our common tasks is to detect and filter these outliers. The main reason for this detection is that the presence of such outliers can cause serious issues in statistical analysis.

EDA

If we want to explain in simple terms, it means trying to understand the given data much better, so that we can make some sense out of it. we are using univariate analysis to describe key characteristics of each feature including, minimum and maximum value, average, standard deviation and others. It was also used to produce a value distribution and identify missing values and outliers.

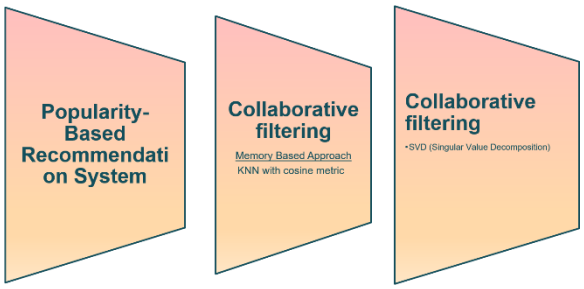
GRAPHICAL REPRESENTATION OF THE RESULTS:

This step involves presenting the dataset with respect to the target feature in the form of graphs, summary tables, Bar chart, Scatter plot, Area plot, stacked plot Pie chart, Table chart, Polar chart, Histogram etc.

EDA CONCLUSIONS:

- The Lovely Bones: A Novel and Wild Animus are the two most read books.
- Most popular book author based on the number of ratings is Stephan King .
- Ballantine Books and Pocket are the top publishers based on the number of ratings that their books have received.
- The majority of readers are between the ages of 20 and 40.
- The majority of readers who have given the books ratings are from the United States and Canada.
- Regardless of the age group, The Lovely Bones and Wild animus appear on lists of the top-rated books.

Recommender System:



1. Popularity Based Recommender System:

It is a type of recommendation system that bases choices on factors like popularity and/or current trends.

	isbn	book-rating	book-title	book-author	year-of-publication	publisher
0	0316666343	707	The Lovely Bones: A Novel	ALICE SEBOLD	2002.0	Little, Brown
1	0971880107	581	Wild Animus	RICH SHAPERO	2004.0	Too Far
2	0385504209	487	The Da Vinci Code	DAN BROWN	2003.0	Doubleday
3	0312195516	383	The Red Tent (Bestselling Backlist)	ANITA DIAMANT	1998.0	Picador USA
4	0060928336	320	Divine Secrets of the Ya-Ya Sisterhood: A Novel	REBECCA WELLS	1997.0	Perennial
5	059035342X	313	Harry Potter and the Sorcerer's Stone (Harry Potter (Paperback))	J. K. ROWLING	1999.0	Arthur A. Levine Books
6	0142001740	307	The Secret Life of Bees	SUE MONK KIDD	2003.0	Penguin Books
7	0446672211	295	Where the Heart Is (Oprah's Book Club (Paperback))	BILLIE LETTS	1998.0	Warner Books
8	044023722X	281	A Painted House	JOHN GRISHAM	2001.0	Dell Publishing Company
9	0452282152	278	Girl with a Pearl Earring	TRACY CHEVALIER	2001.0	Plume Books

2. Weighted average rating method:

Using Weighted average for each Book's Average Rating

W = (Rv + Cm)/(v + m)

where

W= Weighted Rating

R = Average of the Books rating

v = No of people who have rated the books(number of votes)

m = minimum no of votes to be listed

C = the mean rating across all the books

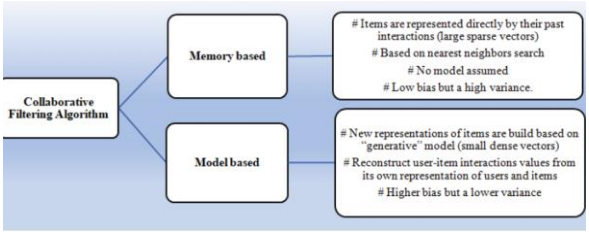
.

	Book-Title	Book-Author	avg_rating	ratings_count	weighted_average
46516	Harry Potter and the Chamber of Secrets Postcard Book	J. K. ROWLING	9.869565	23	9.52
122145	The Two Towers (The Lord of the Rings, Part 2)	J. R. R. TOLKIEN	9.653846	52	9.50
30142	Dilbert: A Book of Postcards	SCOTT ADAMS	9.923077	13	9.36
81784	Postmarked Yesteryear: 30 Rare Holiday Postcards	PAMELA E. APKARIAN-RUSSELL	10.000000	11	9.34
118127	The Return of the King (The Lord of the Rings, Part 3)	J.R.R. TOLKIEN	9.397436	78	9.31
17713	Calvin and Hobbes	BILL WATTERSON	9.583333	24	9.29
100902	The Authoritative Calvin and Hobbes (Calvin and Hobbes)	BILL WATTERSON	9.600000	20	9.25
72637	My Sister's Keeper : A Novel (Picoult, Jodi)	JODI PICOULT	9.545455	22	9.23
118123	The Return of the King (The Lord of The Rings, Part 3)	J. R. R. TOLKIEN	9.625000	16	9.20
120090	The Sneetches and Other Stories	DR. SEUSS	10.000000	8	9.17

3. Collaborative filtering:

- Collaborative Filtering is a Machine Learning technique used to identify relationships between pieces of data. This technique is frequently used in recommender systems to identify similarities between user data and items.

This means that if *Users A* and *B* both like *Product A*, and *User B* also likes *Product B*, then *Product B* could be recommended to *User A* by the system



➢ Memory Based Approach

KNN Based Algorithm:

This algorithm takes into consideration up-to 'K' nearest users (in user based collaborative filtering) or 'K' nearest items (in item based collaborative filtering) for making recommendations. By default, the algorithm is 'user-based', and k is 40 (km in is 1). This means ratings of 40 nearest users are considered while recommending an item to a user. Some variants of this algorithm include with Means, with Zscore & Baseline wherein the average rating of users, or the normalized ZScore of ratings or the baseline rating are also considered as the system generates recommendations

Harry Potter and the Prisoner of Azkaban (Book 3)
Harry Potter and the Goblet of Fire (Book 4)
Harry Potter and the Sorcerer's Stone (Book 1)
Dr. Seuss's A B C (I Can Read It All by Myself Beginner Books)
The Second Generation
Lover Beware
J. K. Rowling: The Wizard Behind Harry Potter
A Dash of Death
So Much to Tell You

➢ Model Based Approach

SVD (Singular Value Decomposition):

This algorithm takes a matrix factorization approach. The user-item rating matrix is factorized into smaller dimension user & item matrices consisting of latent factors (hidden characteristics). By default, number of latent factors is 100. These latent factors are able to capture the known user-item rating preference & in the process are able to predict an estimated rating for all user-item pair where user has not yet rated an item

Model Evaluation:

In Recommender Systems, there are a set metrics commonly used for evaluation. We choose to work with Top-N accuracy metrics, which evaluates the accuracy of the top recommendations provided to a user, comparing to the items the user has actually interacted in test set. This evaluation method works as follows:

- For each user
- For each item the user has interacted in test set
- Sample 100 other items the user has never interacted.
- Ask the recommender model to produce a ranked list of recommended items, from a set composed of one interacted item and the 100 non-interacted items
- Compute the Top-N accuracy metrics for this user and interacted item from the recommendations ranked list
- Aggregate the global Top-N accuracy metrics

	hits@5_count	hits@10_count	interacted_count	recall@5	recall@10	_person_id
36	35	89	545	0.064220	0.163303	11676
202	62	81	139	0.446043	0.582734	98391
271	27	35	93	0.290323	0.376344	153662
60	22	36	88	0.250000	0.409091	16795
474	16	28	73	0.219178	0.383562	95359
485	52	60	72	0.722222	0.833333	114368
390	32	32	61	0.524590	0.524590	104636
456	17	24	54	0.314815	0.444444	158295
660	34	43	54	0.629630	0.796296	123883
659	7	13	53	0.132075	0.245283	35859

CONCLUSIONS:

- 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.
- We evaluated the performance of Singular Value Decomposition based recommender and obtained a Global Recall@5 of 30 % and Recall@10 of 41%

REFERENCES:

- Alma Better Recorded Classes
- wikipedia.org
- analyticsvidhya.com
- thecleverprogrammer.com
- towardsdatascience.com