Book Recommendation Dataset

Background

The last few decades have witnessed the ascendance of platforms such as YouTube, Amazon, Netflix, and various other web services, solidifying the pervasive role of recommender systems in our daily lives. From shaping the landscape of e-commerce by suggesting items that align with buyer preferences to refining online advertising by delivering content tailored to individual user tastes, recommender systems have become an indispensable element of our digital experiences.

In a broad context, recommender systems are algorithms crafted to recommend pertinent items to users, encompassing movies, texts, products, and more, depending on the industry. Their critical role is underscored in industries where their effectiveness can translate into substantial revenue or provide a significant competitive advantage.

A compelling illustration of their importance is the "Netflix prize" from a few years ago, where Netflix organized a challenge inviting participants to develop a recommender system outperforming the platform's own algorithm. The incentive was a remarkable 1 million dollars, highlighting recommender systems' pivotal role in shaping user experiences and industry landscapes.

Objective

Provide traditional book publishing companies with a recommender system to effectively compete and maintain relevance in today's dynamic book market, by leveraging data-driven personalization. This entails implementing a feedback loop within the recommender system, enabling users to rate and provide feedback on suggested books. The goal is to utilize this valuable user data for a continual enhancement of recommendations, ensuring adaptability to the changing tastes and preferences of readers over time.

Scope

In my role as a newly hired Data Analyst at the fictitious Data Analytics company BookWise Metrics, our focus involves analyzing and strategizing for the integration of a new user feedback system.

Data Source

The Book Recommendation Dataset is available for access on Kaggle.

The owner of the dataset is Arash Nicoomanesh (Möbius), a Senior Data Scientist from Iran and a KaggleX BIPOC Mentor and Kaggle Master.

The data was compiled by Cai-Nicolas Ziegler in a 4-week crawl (August/September 2004) from the Book-Crossing community with kind permission from Ron Hornbaker, CTO of Humankind Systems.

This dataset was uploaded to Kaggle on February 9th 2024.

Relevance

I am confident that this dataset fulfills the project criteria effectively, being open source, incorporating a geospatial component, and meeting the specified size and variable requirements. It's noteworthy that the dataset is only three weeks old. I selected this dataset based on its alignment with the outlined project brief requirements, and it holds the potential for creating a compelling and insightful project within the book industry, which is the sector I aspire to work in.

Data Profile

Data Collection

The data originates from the Book-Crossing community, scraped with permission by Cai-Nicolas Ziegler.

Content

The Book-Crossing dataset comprises 3 files:

- **Users**
  - Contains the users. Note that user IDs (User-ID) have been anonymized and mapped to integers. Demographic data is provided (Location, Age) if available. Otherwise, these fields contain NULL values
- **Books**
  - Books are identified by their respective ISBNs. 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, and 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

Variable	Time-variant/ Time-invariant	Structured/ Unstructured	Qualitative/ Quantitative	Qualitative: Binary/Nominal/Ordinal Quantitative: Discrete/Continuous
ISBN	invariant	structured	qualitative	nominal
bookTitle	invariant	structured	qualitative	nominal
bookAuthor	invariant	structured	qualitative	nominal
yearOfPublication	invariant	structured	quantitative	discrete
publisher	invariant	structured	qualitative	nominal
imageUrlS	invariant	unstructured	qualitative	ordinal
imageUrlM	invariant	unstructured	qualitative	ordinal
imageUrlL	invariant	unstructured	qualitative	ordinal
User-ID	invariant	structured	quantitative	discrete
Location	variant	structured	qualitative	nominal
Age	variant	structured	quantitative	discrete
Book-Rating	variant	structured	quantitative	discrete

Data Limitations

The data is static and won't undergo updates. Due to the absence of integrated revenue data, making financial predictions is not feasible. A thorough review of the columns, followed by data cleaning, is imperative to ensure the correct format before commencing the analysis.

Data Ethics

The dataset contains 278,858 users (anonymized but with demographic information) providing 1,149,780 ratings (explicit/implicit) about 271,379 books, thus respecting privacy and being open source.

Data Wrangling

Column(s) renamed	Column(s) dropped	Comments
Book-Title		To remove the hyphen(s) and have all columns starting with a lowercase for consistency
Book-Author		To remove the hyphen(s) and have all columns starting with a lowercase for consistency
Year-Of-Publication		To remove the hyphen(s) and have all columns starting with a lowercase for consistency
Publisher		To remove the hyphen(s) and have all columns starting with a lowercase for consistency
Image-URL-S		To remove the hyphen(s) and have all columns starting with a lowercase for consistency
Image-URL-M		To remove the hyphen(s) and have all columns starting with a lowercase for consistency
Image-URL-L		To remove the hyphen(s) and have all columns starting with a lowercase for consistency
	imageUrlS	Not be required for the analysis
	imageUrlM	Not be required for the analysis
	imageUrlL	Not be required for the analysis

Column(s) type changed	Column(s) dimension checking	Comments
BookAuthor		Mixed value data type found and updated (ref. Data Cleaning section)
YearofPublication		Changed all values as integers, setting invalid years as NaN and replacing them with the mean value
	YearofPublication	bookAuthor is incorrectly loaded with bookTitle > updated
	publisher	Checking rows with NaN to find a pattern > no clues found, NaN replaced by 'other'
	Location	Split the location column into City and Country

Data Cleaning

Missing Values	Duplicates	Mixed Data Types
None	None	BookAuthor > type changed to a string 'str'

Data Merge

The data frames were merged in 2 steps:

1. The books and ratings df were joined with the 'ISBN' variable as 'books_ratings'
2. The new 'books_ratings' was merged with the users' df with the 'User-ID' variable as 'books_ratings_users'

Statistical Summary

```
books_ratings_users.describe()
```

	yearOfPublication	User-ID	Book-Rating	Age
count	1.031136e+06	1.149780e+06	1.149780e+06	1.149780e+06
mean	1.995283e+03	1.403864e+05	2.866950e+00	3.608129e+01
std	7.309099e+00	8.056228e+04	3.854184e+00	1.032642e+01
min	1.376000e+03	2.000000e+00	0.000000e+00	5.000000e+00
25%	1.992000e+03	7.034500e+04	0.000000e+00	3.100000e+01
50%	1.997000e+03	1.410100e+05	0.000000e+00	3.400000e+01
75%	2.001000e+03	2.110280e+05	7.000000e+00	4.100000e+01
max	2.006000e+03	2.788540e+05	1.000000e+01	9.000000e+01

Questions

1. What geographical cities and/or countries show the highest and lowest ratings?
2. Are there any notable patterns in reading levels based on location?
3. Which books are most popular among users in the dataset?
4. Can we identify trends in the popularity of specific authors or titles over time?
5. Are there noticeable differences in reading levels among age groups?

Acknowledgements

A very special thank you to the following people for making the data available to the public!!!

- Arash Nicoomanesh (Möbius) for creating the dataset
- Cai-Nicolas Ziegler for collecting the data from the Book-Crossing community with kind permission from Ron Hornbaker, CTO of Humankind Systems