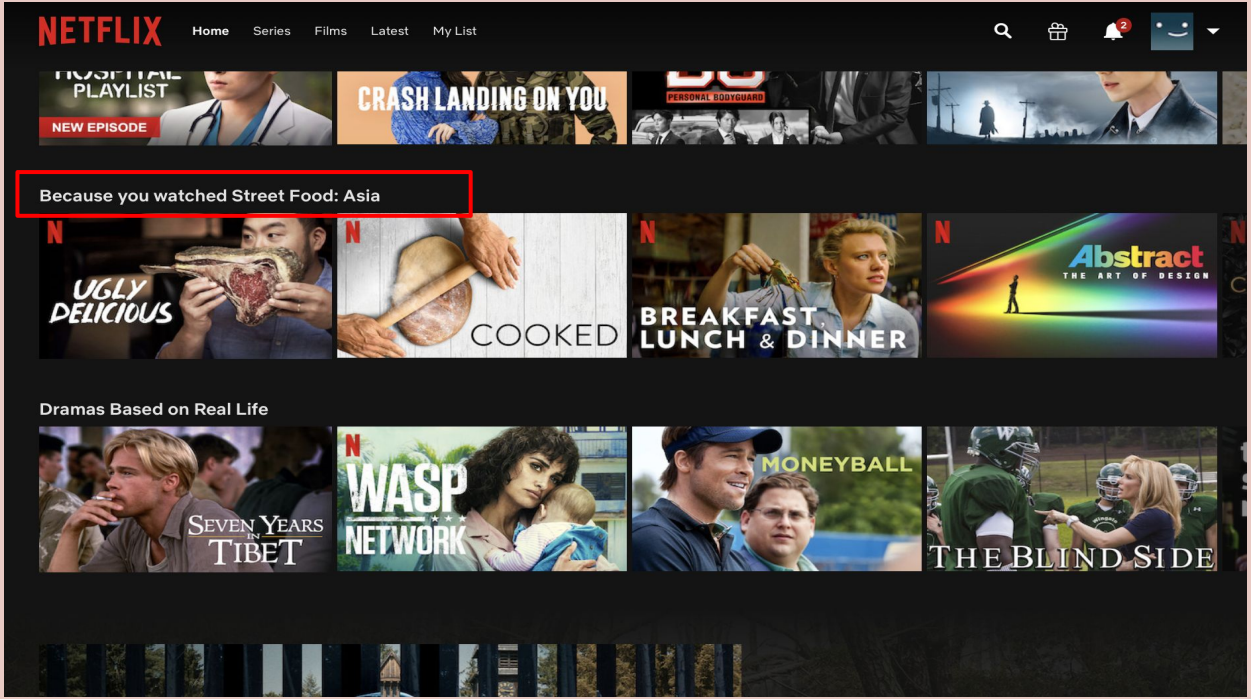
Book Recommendation System

DSI14 Capstone

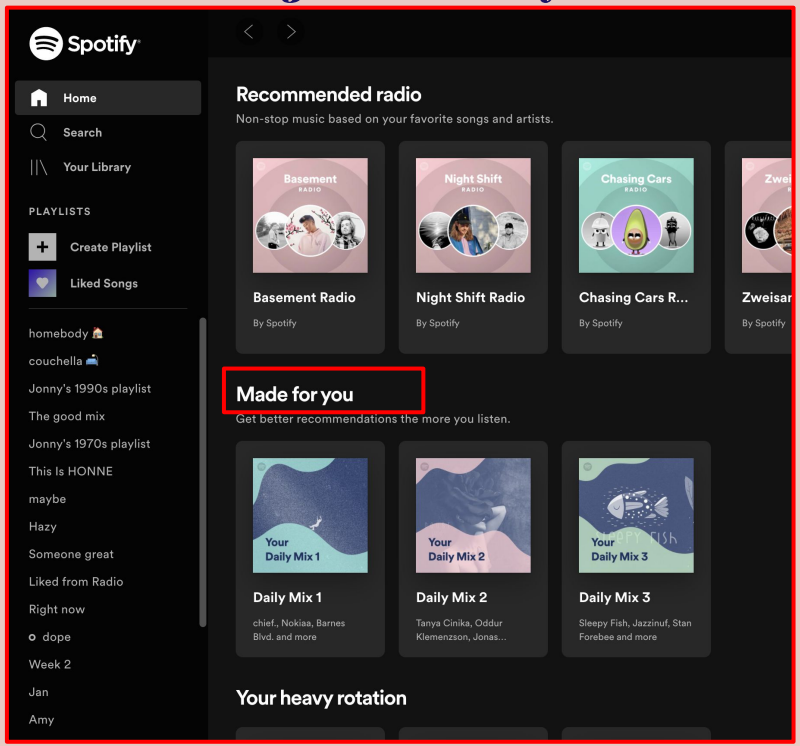
Jocelyn Pok

Table of Contents

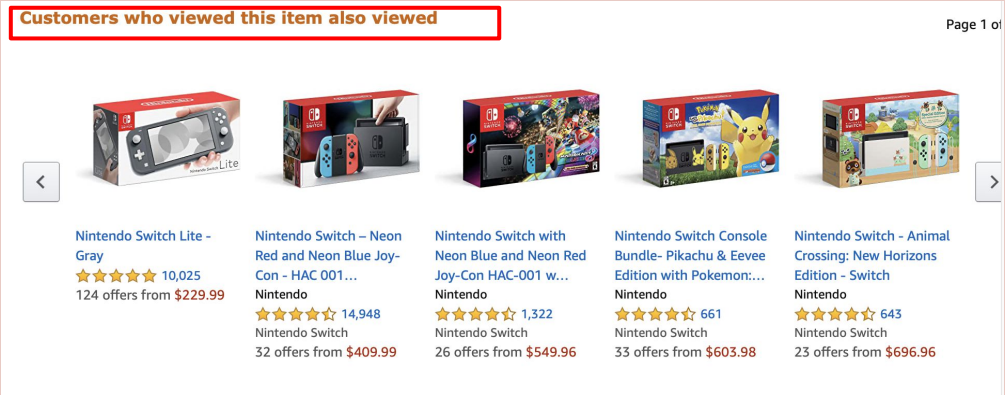
Use case 1 : User-item
“ Recommended just for you”



Use case 1 : User-item
“ Recommended just for you”



Use case 2 : Item-item
“ Similar products...”



Amazon

According to Mckinsey & Company, 35% of Amazon's revenue is generated by it's recommendation engine

Netflix

According to Mckinsey & Company, 75% of what users watch on Netflix come from product recommendations

Spotify

Possibly one of Spotify's most innovative uses of AI and recommendation systems is their popular Discover Weekly playlist. Known as Release Radar, this algorithmically powered tool updates personal playlists on a weekly basis so that users won't miss newly released music by artists they like.

The new recommendation system has helped Spotify increase its number of monthly users from 75 million to 100 million at a time, in spite of competition from rival streaming service Apple Music.

Situation

The New York Times

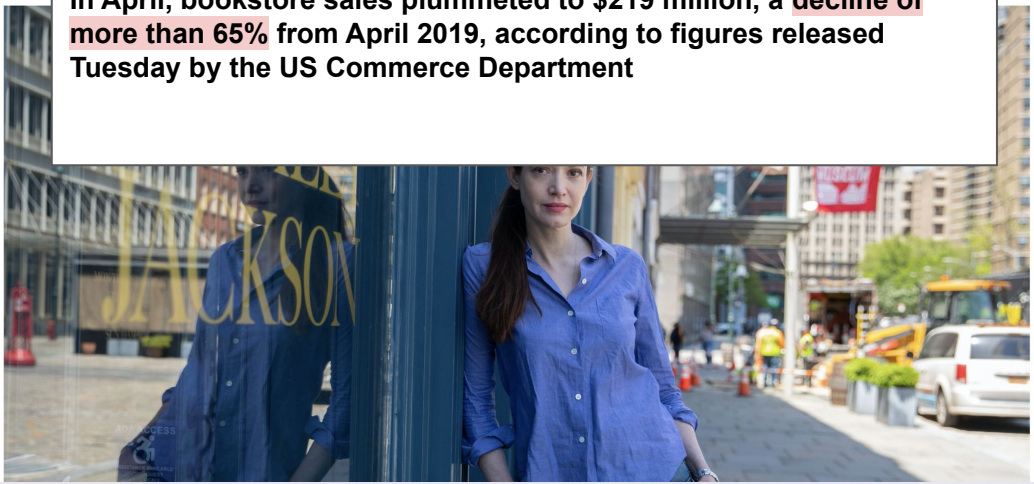
SUBSCRIBE NOW

Opinion

What Could Kill My New York Bookstores?

It won't be Amazon or the coronavirus. It will be artificially high rents.

In April, bookstore sales plummeted to $219 million, a decline of more than 65% from April 2019, according to figures released Tuesday by the US Commerce Department



Problem Statement

“COVID-19 is not only a public health issue. It is also a serious economic, social and political problem” -PM Lee

Beyond the pandemic, independent bookstores that have joined the online retail ecosystem can explore implementing recommendation systems that can significantly boost sales and positive effects on the user experience as well, which translates into metrics that are of much importance to online businesses such as customer satisfaction and retention.

This project seeks to identify the best recommendation model that obtains the best top-N accuracy metrics for the dataset. The methods that will be explored in this project include : content-based, collaborative filtering and hybrid

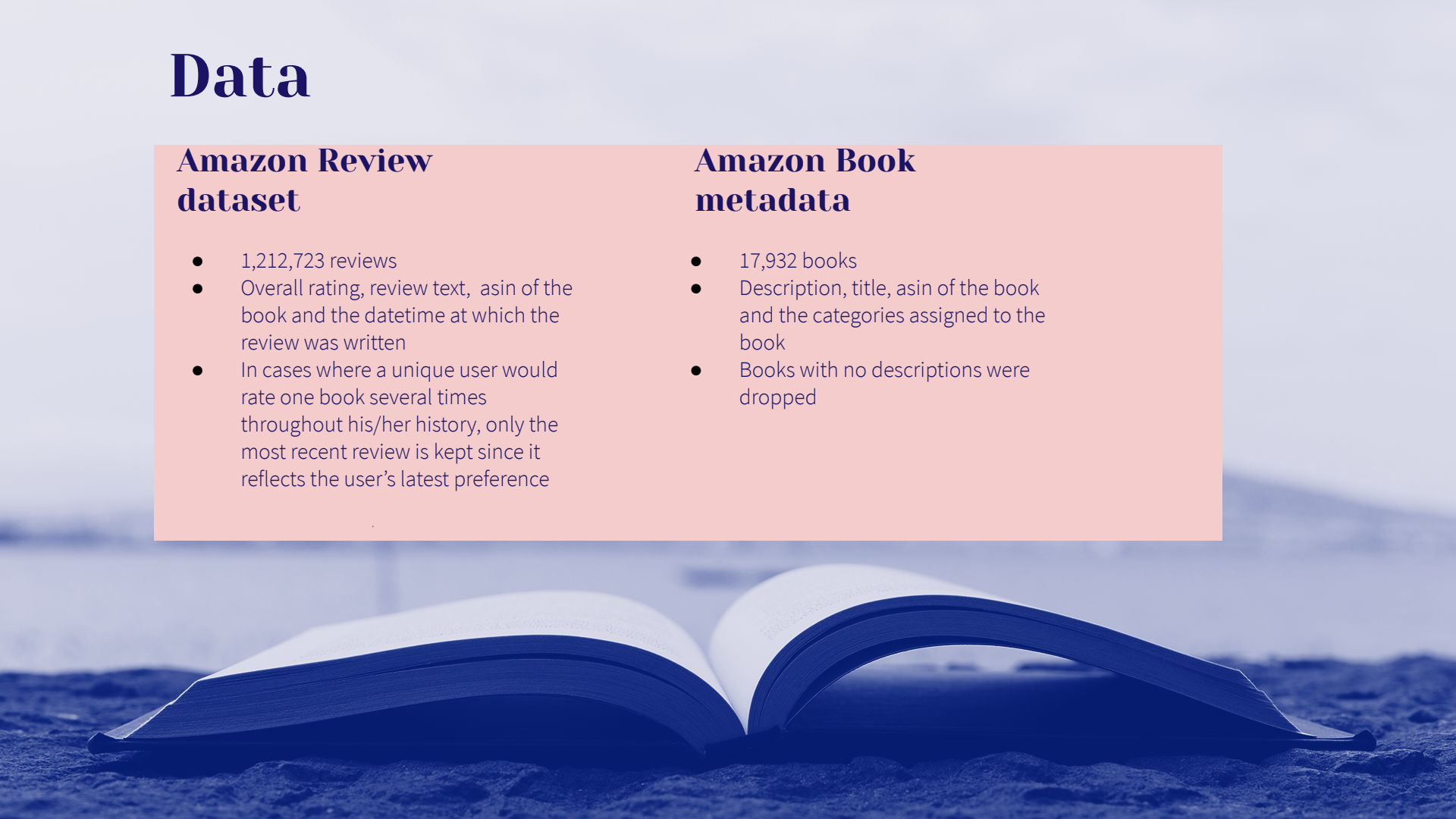
Data

Amazon Review dataset

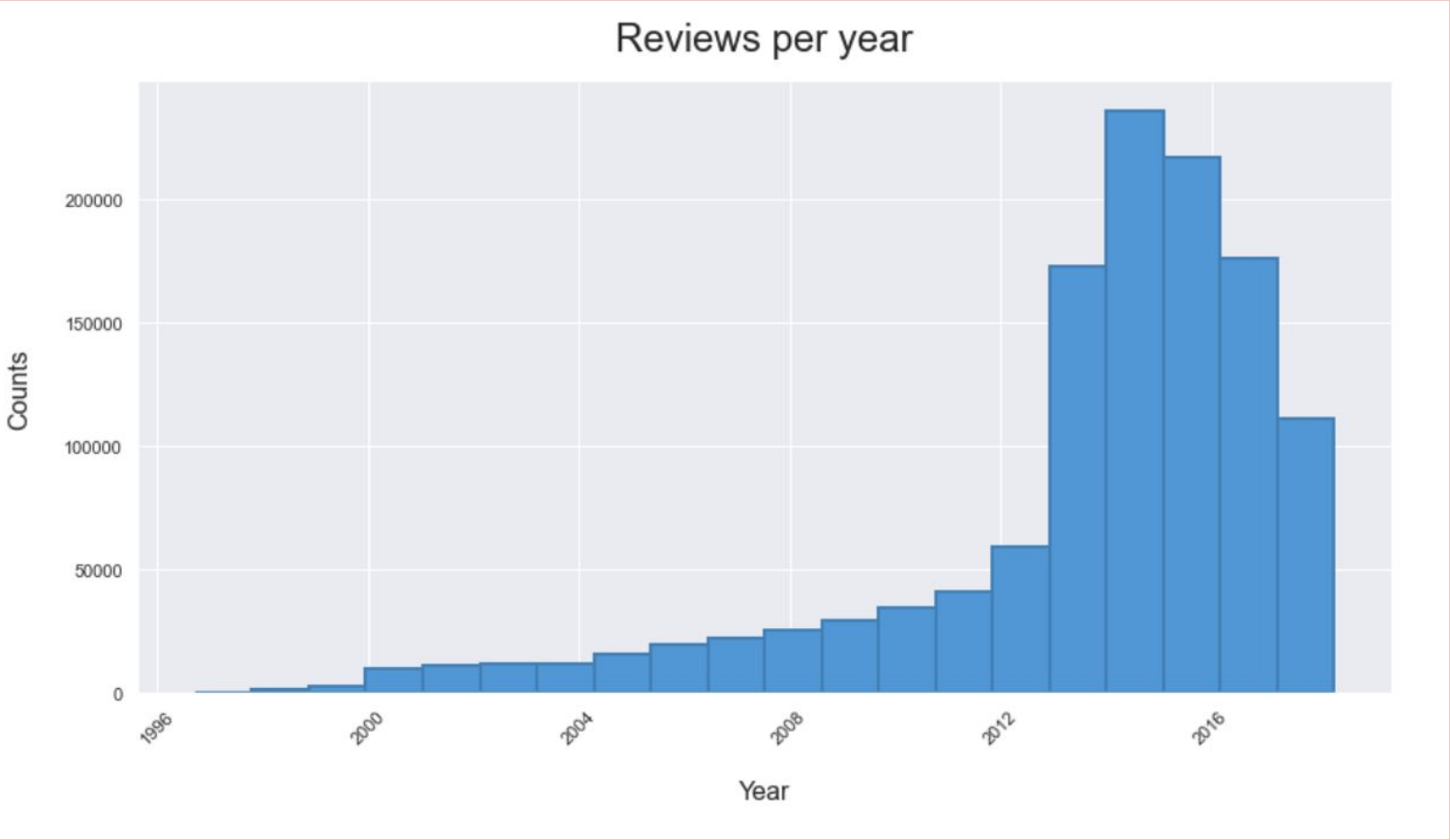
- 1,212,723 reviews
- Overall rating, review text, asin of the book and the datetime at which the review was written
- In cases where a unique user would rate one book several times throughout his/her history, only the most recent review is kept since it reflects the user's latest preference

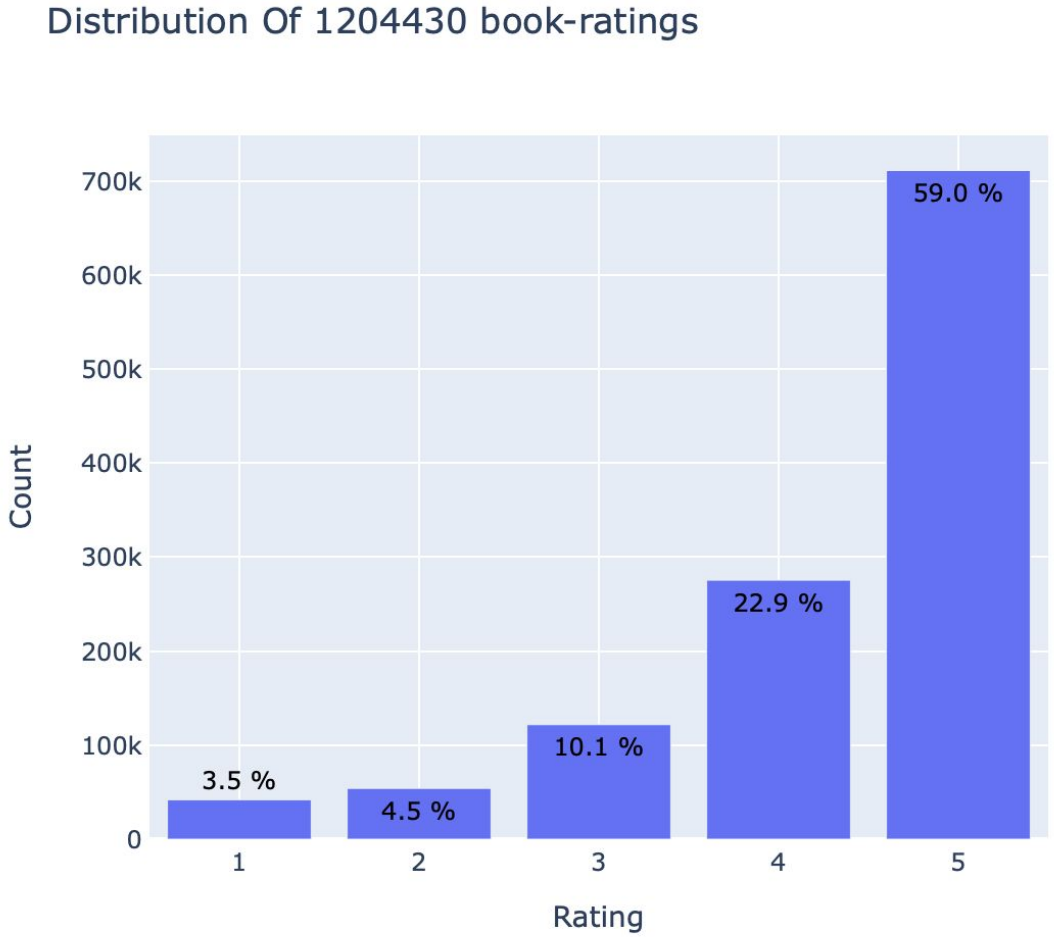
Amazon Book metadata

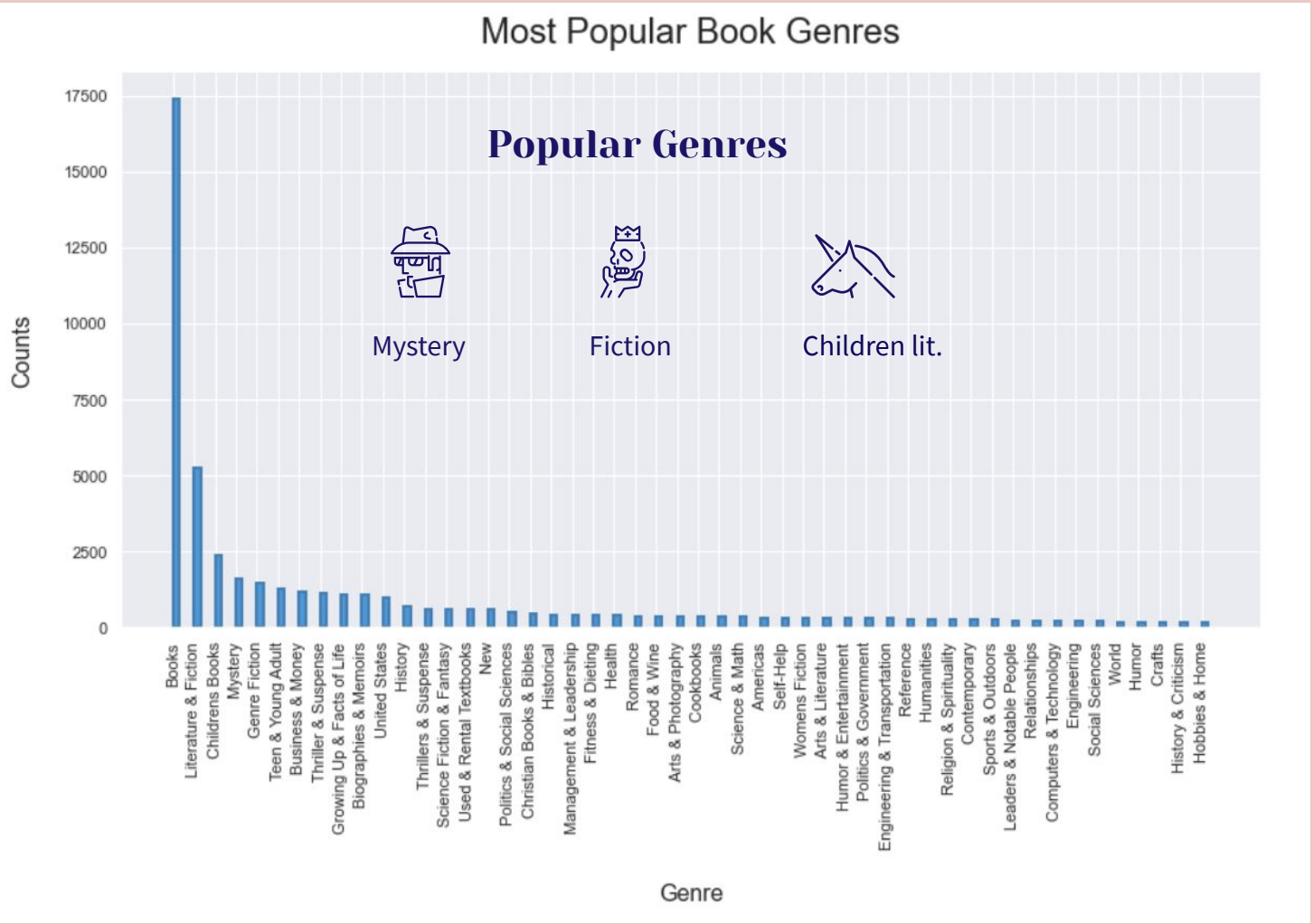
- 17,932 books
- Description, title, asin of the book and the categories assigned to the book
- Books with no descriptions were dropped



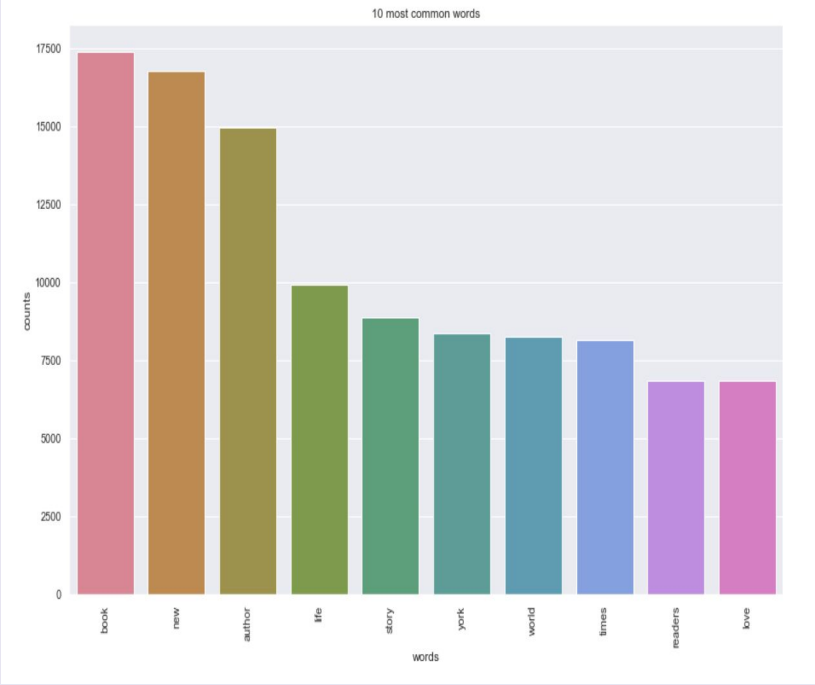
Exploratory Data Analysis


Reviews per year
Counts
200000
150000
100000
50000
0
1996
2000
2004
2008
2012
2016
Year

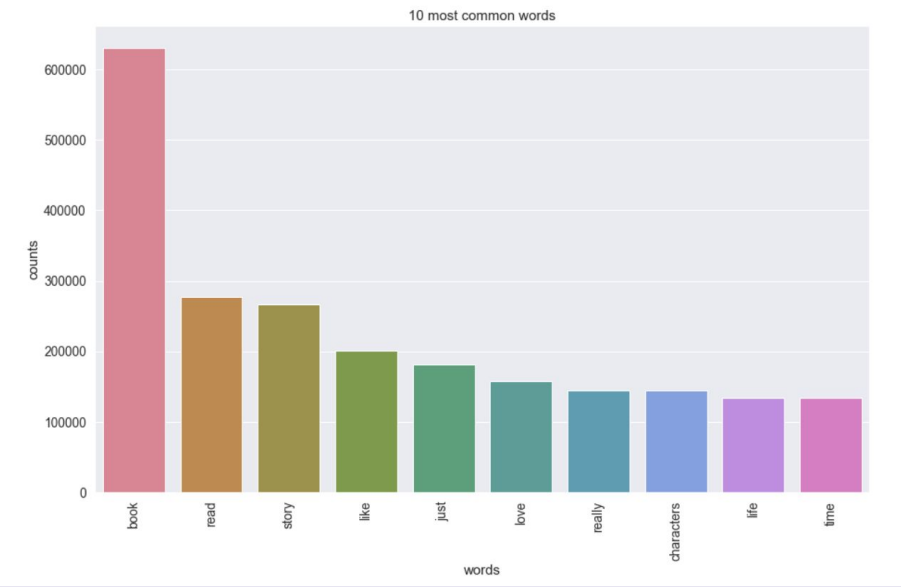

Distribution Of 1204430 book-ratings
Count
700k
600k
500k
400k
300k
200k
100k
0
3.5 %
4.5 %
10.1 %
22.9 %
59.0 %
1
2
3
4
5
Rating

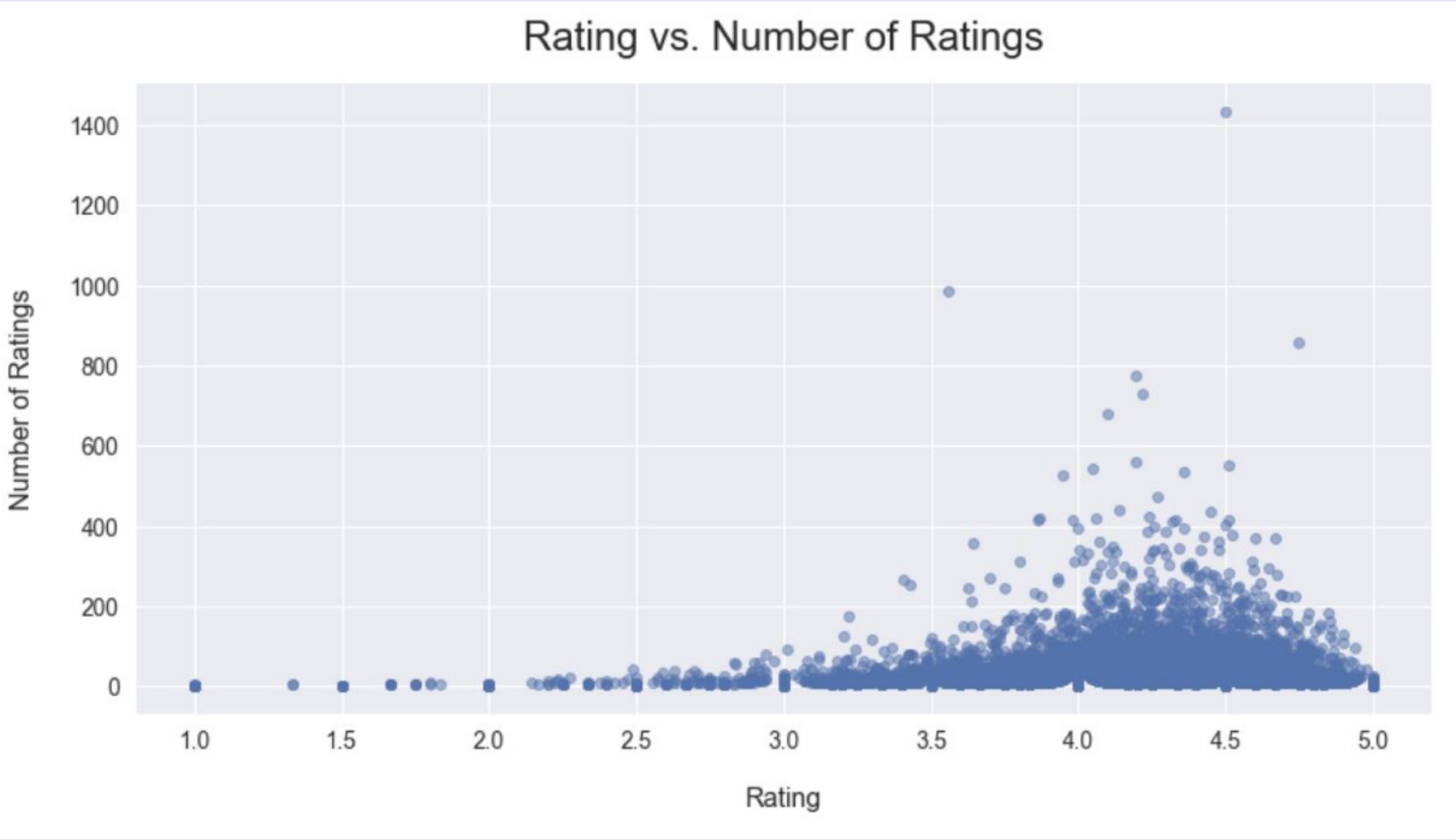

Most Popular Book Genres
Popular Genres
Mystery
Fiction
Children lit.
Counts
17500
15000
12500
10000
7500
5000
2500
0
Books
Literature & Fiction
Childrens Books
Mystery
Genre Fiction
Teen & Young Adult
Business & Money
Thriller & Suspense
Growing Up & Facts of Life
Biographies & Memoirs
United States
History
Thrillers & Suspense
Science Fiction & Fantasy
Used & Rental Textbooks
New
Politics & Social Sciences
Christian Books & Bibles
Historical
Management & Leadership
Fitness & Dieting
Health
Romance
Food & Wine
Arts & Photography
Cookbooks
Animals
Science & Math
Americas
Self-Help
Womens Fiction
Arts & Literature
Humor & Entertainment
Politics & Government
Engineering & Transportation
Reference
Humanities
Religion & Spirituality
Contemporary
Sports & Outdoors
Leaders & Notable People
Relationships
Computers & Technology
Engineering
Social Sciences
World
Humor
Crafts
History & Criticism
Hobbies & Home
Genre

Description



Reviews




Rating vs. Number of Ratings
Number of Ratings
1400
1200
1000
800
600
400
200
0
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
Rating

Model fit and Evaluation

Measure model performance

Top N-accuracy

- Users are provided with a ranked list of N items they will be interested in, in order to encourage views and purchase

Recall@N

- Hit@N count/ total interacted items in test set

Hit Rate

- To evaluate the top-N recommendations, if the user has rated one of the top-N recommended items, it is considered a “hit”

Baseline Approach

“In the recommendations world, there’s a cardinal rule. If i know nothing about you, then the best things to recommend to you are the most popular things in the world”
- Vijai Mohan, Senior Applied Scientist, Personalization Team of Amazon

Which model to use?

Recommendation algorithms

Collaborative filtering

- Matrix completion
- User/Item similarity

Content-based Models

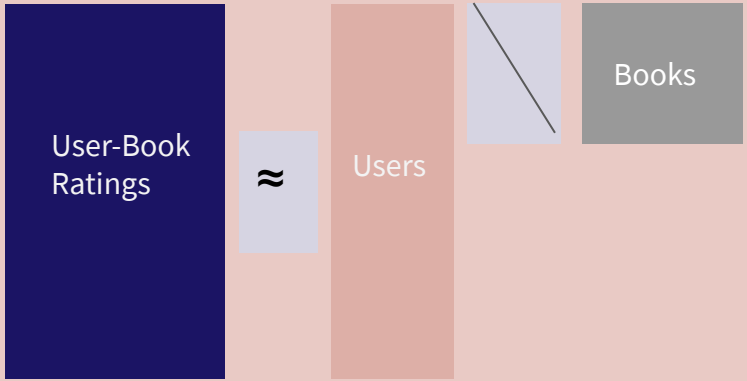
- Match based on user and book features

Hybrid Models

Modelling

Recommendation Algorithms

Collaborative Filtering



- Singular Value Decomposition(SVD)
- Decomposes user-book interactions matrix into a set of user and book vectors
- Use vectors to estimate user responses to unseen items

Content-based Models

Extract features out of the book description to create useful representations

- Vector for each document is represented as the computed TF-IDF weights
- Find similar books by comparing their word vector

Comparison of methods

modelName	recall@5	recall@10
Popularity	0.413076	0.542086
Collaborative Filtering	0.576140	0.710156
Content-Based	0.898594	0.911581
Hybrid	0.542114	0.542443

05

Business Recommendations

Your subtitle goes here

The content-based method is well-suited for businesses new to the realm of recommendation systems but seek to incorporate an element of personalization. By identifying user's specific tastes and then provide them with relevant products will enhance the user's experience.

Limitations

A content-based system is limited by the novelty in recommended results. Because it only looks at user's past history, it never jumps to other areas that users might like but never interacted before.
Additionally, if the quality of the content does not contain enough information to discriminate the items precisely, this method will perform poorly

Looking forward

Explore

- More complex representations of data(eg word embeddings)
- Deep learning: Neural network architectures for collaborative filtering can be explored.

Deployment

- **A/B testing**: Assign users to treatment(new model) and control(existing) groups.
- Evaluate outcomes; CTR ,sales of the books
- Complex models with lots of data take a long time to train
- Do **scheduled** batch prediction
- Save results to **in-memory datastores** for fast retrieval(Redis, Memcached)
- Enable **autoscaling** on API servers

Demo


slidesgo
Free Google Slides and PowerPoint templates to boost your presentations
Most Popular Google Slides themes and Powerpoint templates
Memphis Marketing Plan
Orange Memphis
Back to School Social Media
COMMUNICATION CONSULTING
TECH STARTUP

Thanks