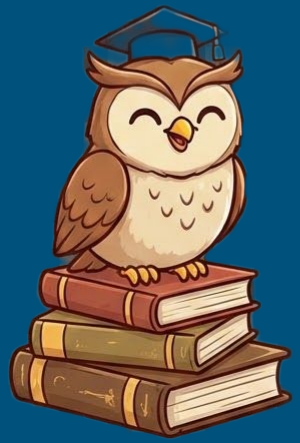
Goodreads Recommendation

Assignment 1: Problem Proposal + Data Exploration



Manan Agarwal
Saksham Singh
Sheel Shah

Problem Proposal

The Goodreads dataset is a large-scale, web-scraped collection from goodreads.com containing:

1. Detailed book metadata
2. User profiles and interaction histories
3. Full review text and engagement signals

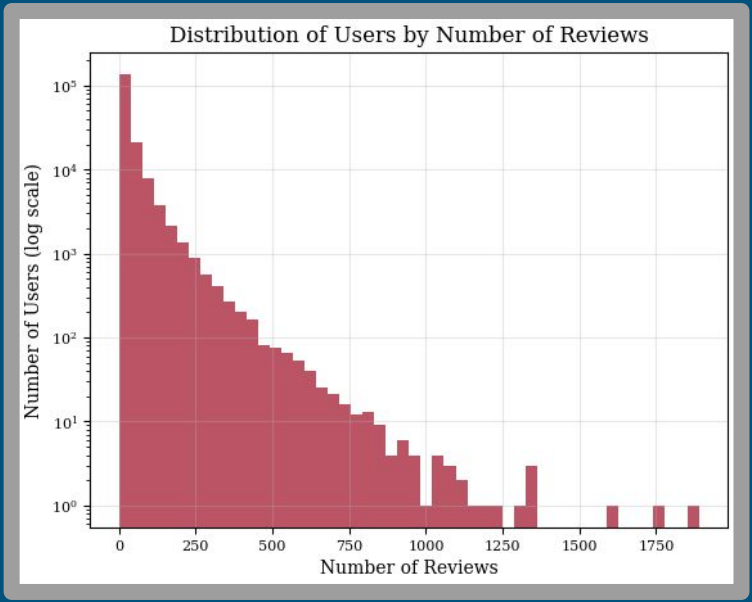
We aim to build a modern recommendation system that leverages:

1. Semantic signals from review text
2. Temporal patterns in user activity
3. Community validation signals (e.g., upvotes)

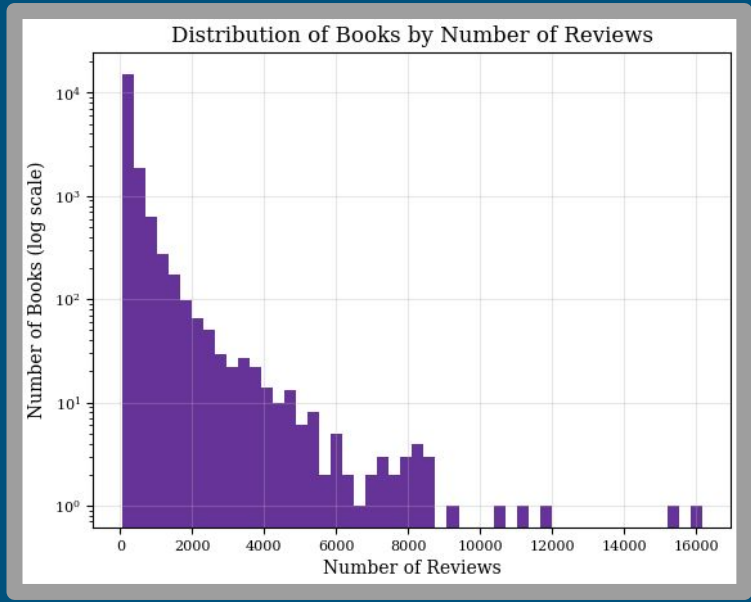
Goodreads Dataset

- **Books:**
  - Book metadata (Title/Author/etc)
  - Genre
  - Average Rating
  - Number of Reviews/ Ratings
- **Reader:**
  - Interactions (read, rating)
  - Review Text
  - Upvotes
  - Reading Duration (start, end)

Preprocessing

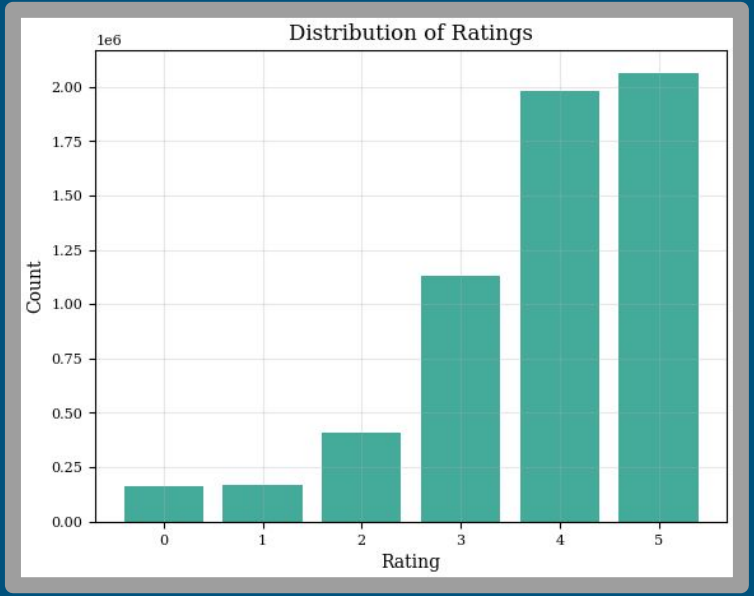


Minimum activity threshold (10 reviews) applied to filter inactive users.



Low-review books (below 100 reviews) are filtered out to avoid noisy aggregate statistics.

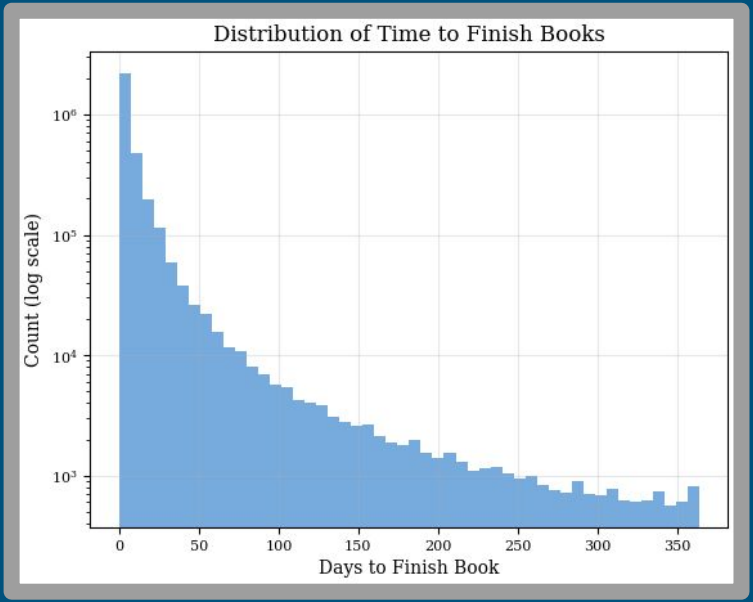
Class Imbalance



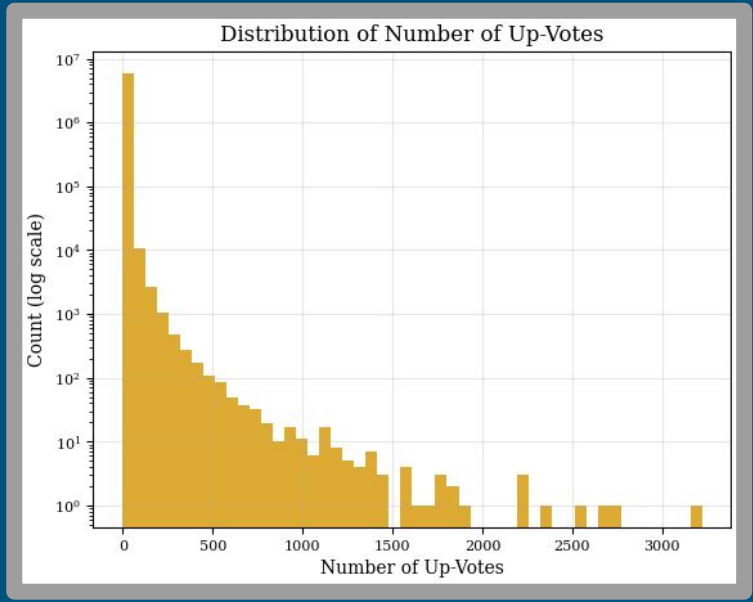
To handle the class-imbalance in ratings, data is subsampled to have equal number of reviews for each rating.

Stage	Users	Books	Reviews
Original	450K	2M	15M
Min-Reviews	200K	20K	5M
Sub-sampled	150K	20K	1M

Review Statistics

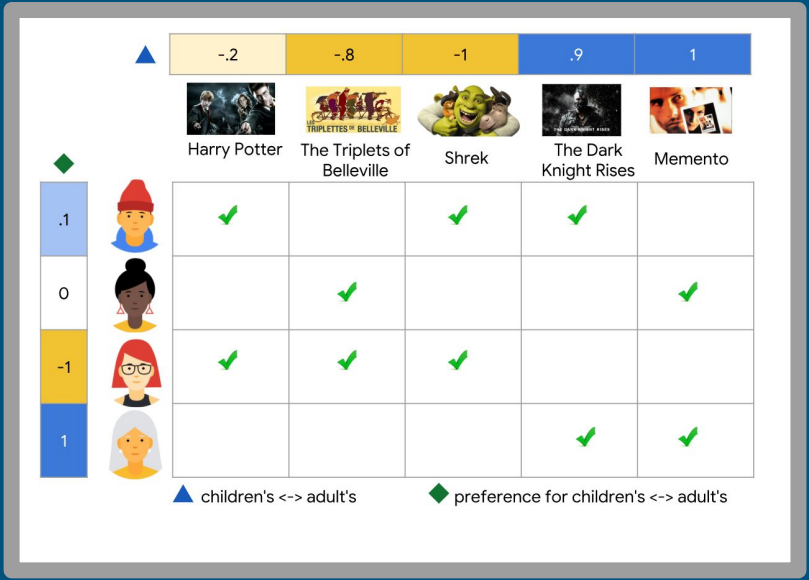


Time-to-completion might serves as a proxy for reader engagement.



Up-votes provide a community-driven signal of perceived review usefulness.

Collaborative Filtering

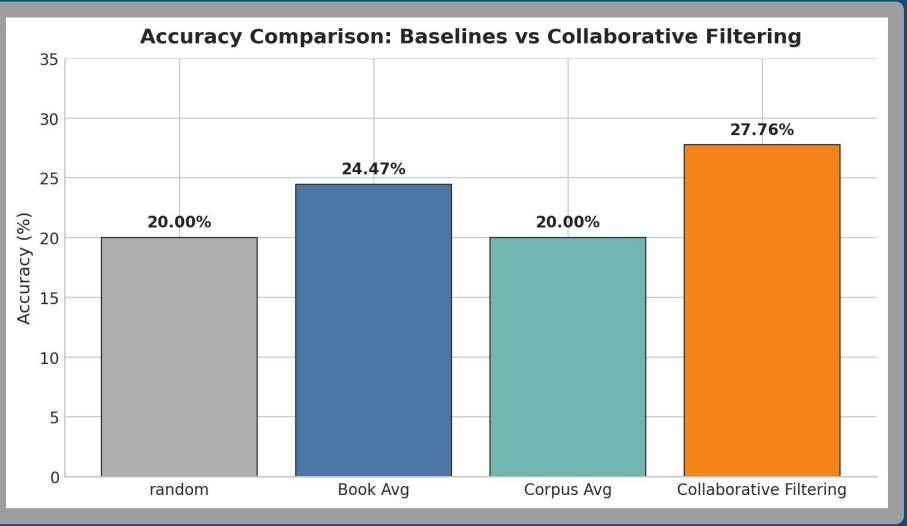


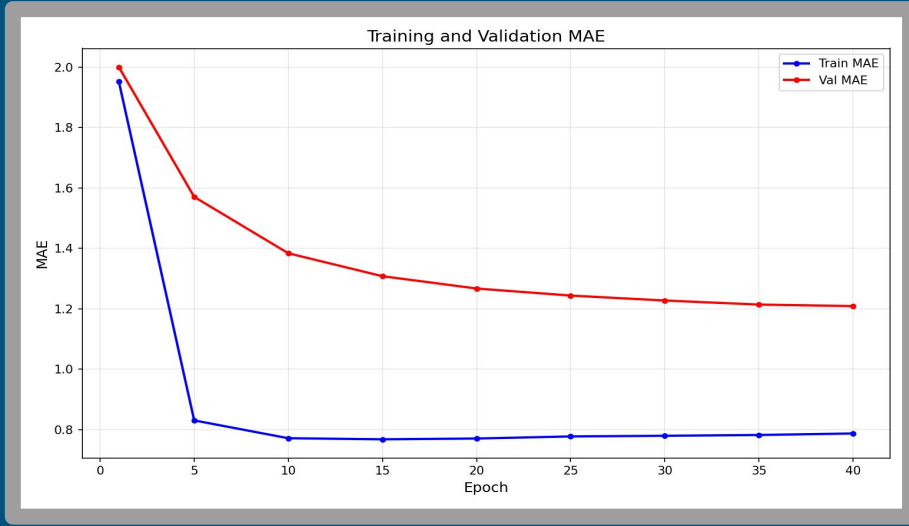
- **Goal** - Predict a rating (1 - 5).
- Collaborative Filtering via Matrix Factorization.

- Baselines
  - Random Guess
  - Predict the Book Average
  - Predict the Dataset Average

- Metrics
  - MAE
  - Accuracy

Evaluation


Accuracy Comparison: Baselines vs Collaborative Filtering
Accuracy (%)
0
5
10
15
20
25
30
35
20.00%
24.47%
20.00%
27.76%
random
Book Avg
Corpus Avg
Collaborative Filtering


Training and Validation MAE
Train MAE
Val MAE
MAE
0.8
1.0
1.2
1.4
1.6
1.8
2.0
0
5
10
15
20
25
30
35
40
Epoch

Method

Limitation of Standard Collaborative Filtering:

1. Traditional models rely only on user–book interaction matrices.
2. They ignore rich behavioral and linguistic signals available on review platforms.

Goal: Augment rating prediction with structured behavioral and semantic features:

1. Semantic Patterns in Prior Reviews: (Topics, Likes and Dislikes)
2. Reading Diversity & Specialization: (Genre Preferences)
3. Reviewer Behavior: (Calibrating Ratings, DNFs, Upvotes)
4. Conformity Effects: (Influence of existing aggregate ratings)