
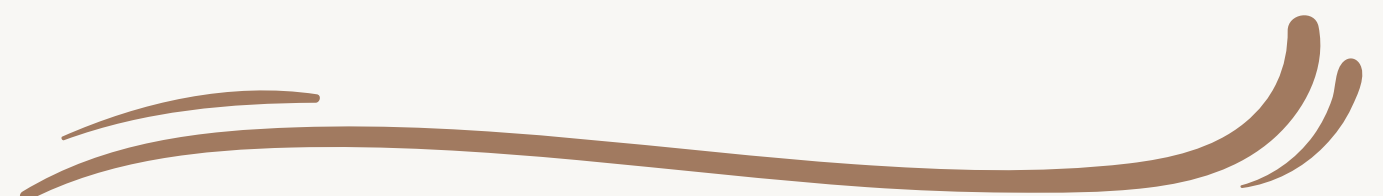




# BOOK RECOMMENDATIONS SYSTEM - PART 2



★ People who love reading  
by Mili Thakrar



# INTRODUCTION

Leverage machine learning to deliver personalized book recommendations, helping readers discover titles that truly resonate with their unique tastes in a sea of literature.

## Problem Area

- Too many books, hard to choose.
- Reduced motivation to read.
- Overwhelmed by recommendations.
- Hard to find niche interests.
- Generic suggestions don't fit.





# DATA SCIENCE SOLUTION


BUILDING A SMART BOOK MATCH MAKER



## Hybrid Recommendation Engine:

- Collaborative + Content-based filtering
- Considers: reading history, ratings, similar readers, book "DNA"

## Process:

- Analyze reader data
  - Decode book essence
  - Develop reader-book matching algorithms
  - Fine-tune recommendations
- 



# WORKFLOW

1

## Data preprocessing and Cleaning

- Analyzing data quality & missing values
- Feature engineering and combining data

2

## Data Analysis and Visualization

- Looking at the distribution and key visualizations
- Hypothesis testing

3

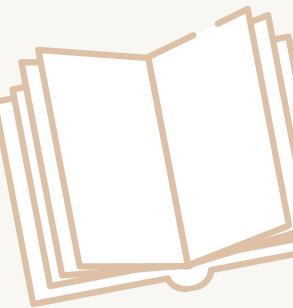
## Baseline Modeling

- Popularity based model
- Baseline Logistic Regression model
- Content based - Cosine Similarity with Word Embedding

4

## Advance Modeling

- Collaborative filtering Model
- Hybrid (Mix of content and collaborative)



# 1 Data Preprocessing

## Prepossessing steps

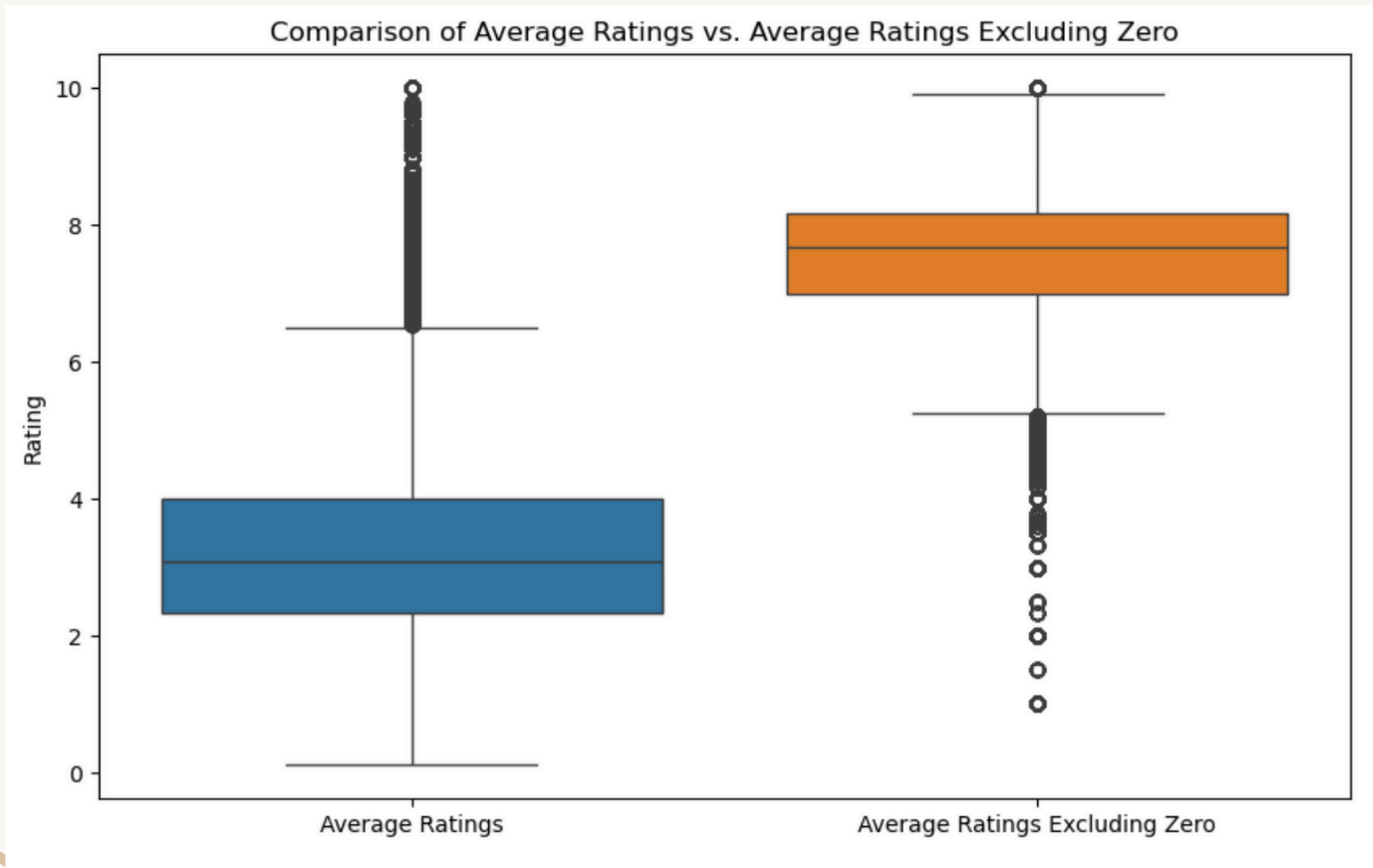
- Accounted for arbitrary values
- combined all three csv files into one
- Feature engineered columns such as Avg. ratings

## Main Dataset

- 851,505 rows and 16 columns
- Target for Baseline - Ratings

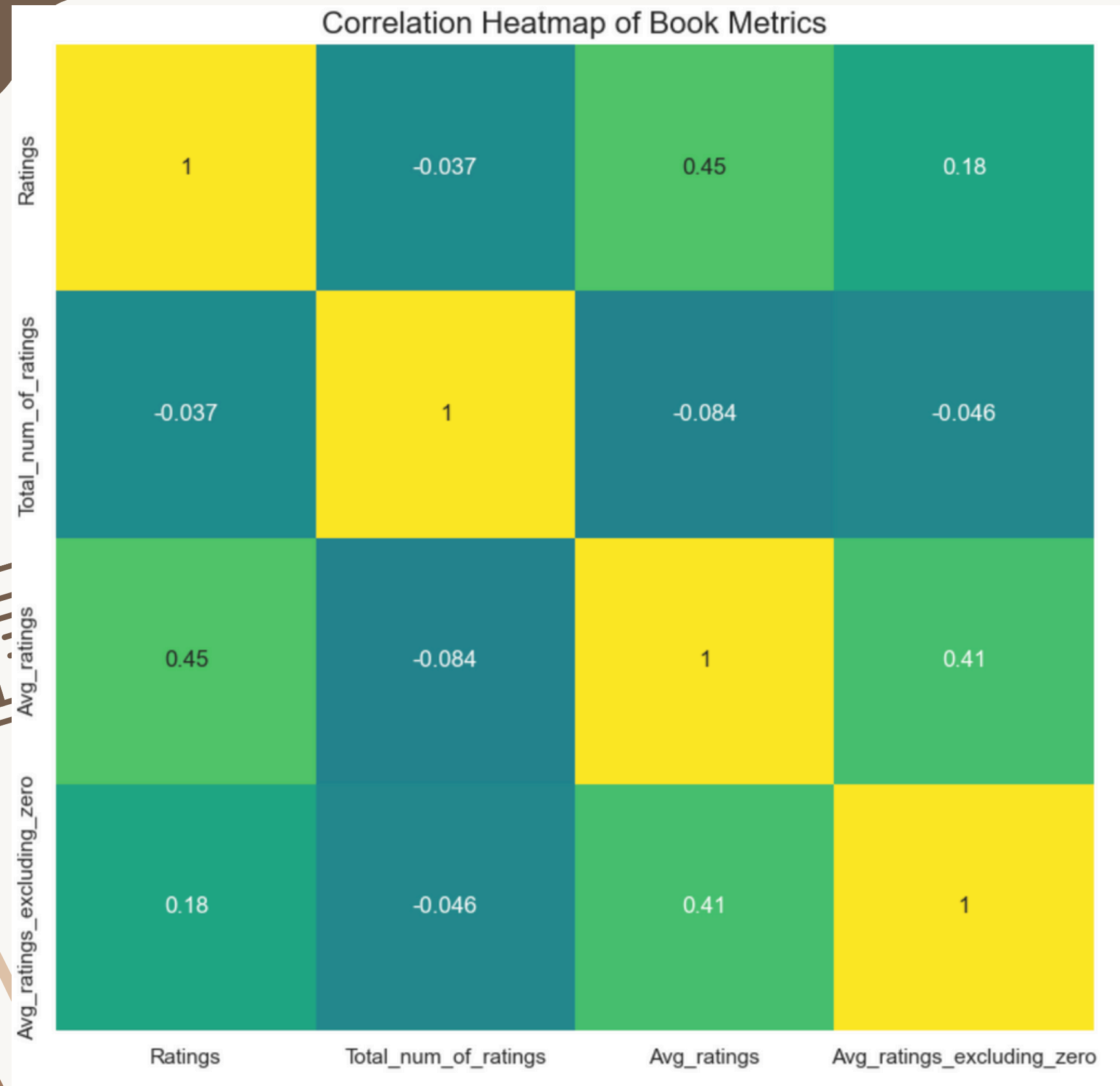


# 2 Data Visualization



- Excluding zero ratings significantly increases the average rating
- Shifts the median from around 3 to about 8.
- The spread of ratings is much wider when zeros are included, while excluding zeros

# Feature Correlation



- **Average Ratings Correlation:** Average ratings have a moderate positive correlation (0.4-0.45) with Ratings and Average ratings excluding zero.
- **Minimal Multicollinearity:** Most features show very low correlation, indicating little multicollinearity.
- **Total Ratings Disconnect:** The total number of ratings has very little correlation with other metrics.

# HYPOTHESIS TESTING

Feature	Correlation	P-Value	Significance
Total_num_of_ratings	-0.1559	0.0000	Significant
Avg_ratings	0.3976	0.0000	Significant
Avg_ratings_excluding_zero	0.1907	0.0000	Significant
Publication_year	-0.0029	0.0064	Significant
Year_Category_2010 onwards	-0.0001	0.9439	Not Significant
Age_Category_26-32	0.0006	0.5495	Not Significant

- Average ratings (including excluding zeros) and total ratings significantly influence review scores.
- Year and age categories weakly but often significantly correlate with review scores

	Chi-Square Statistic	P-Value	Degrees of Freedom
Publisher	2.865993e+05	0.0	115720.0
Year_Category	2.815557e+03	0.0	60.0
Age_Category	3.049193e+03	0.0	50.0
City	5.062891e+05	0.0	142120.0
State	9.669428e+04	0.0	18590.0
Country	2.604930e+04	0.0	3870.0
Title	2.152075e+06	0.0	1355620.0
Author	1.166984e+06	0.0	621090.0

- Categorical features and Ratings are statistically significant.



# 3 Baseline Models

## POPULARITY BASED RANKING

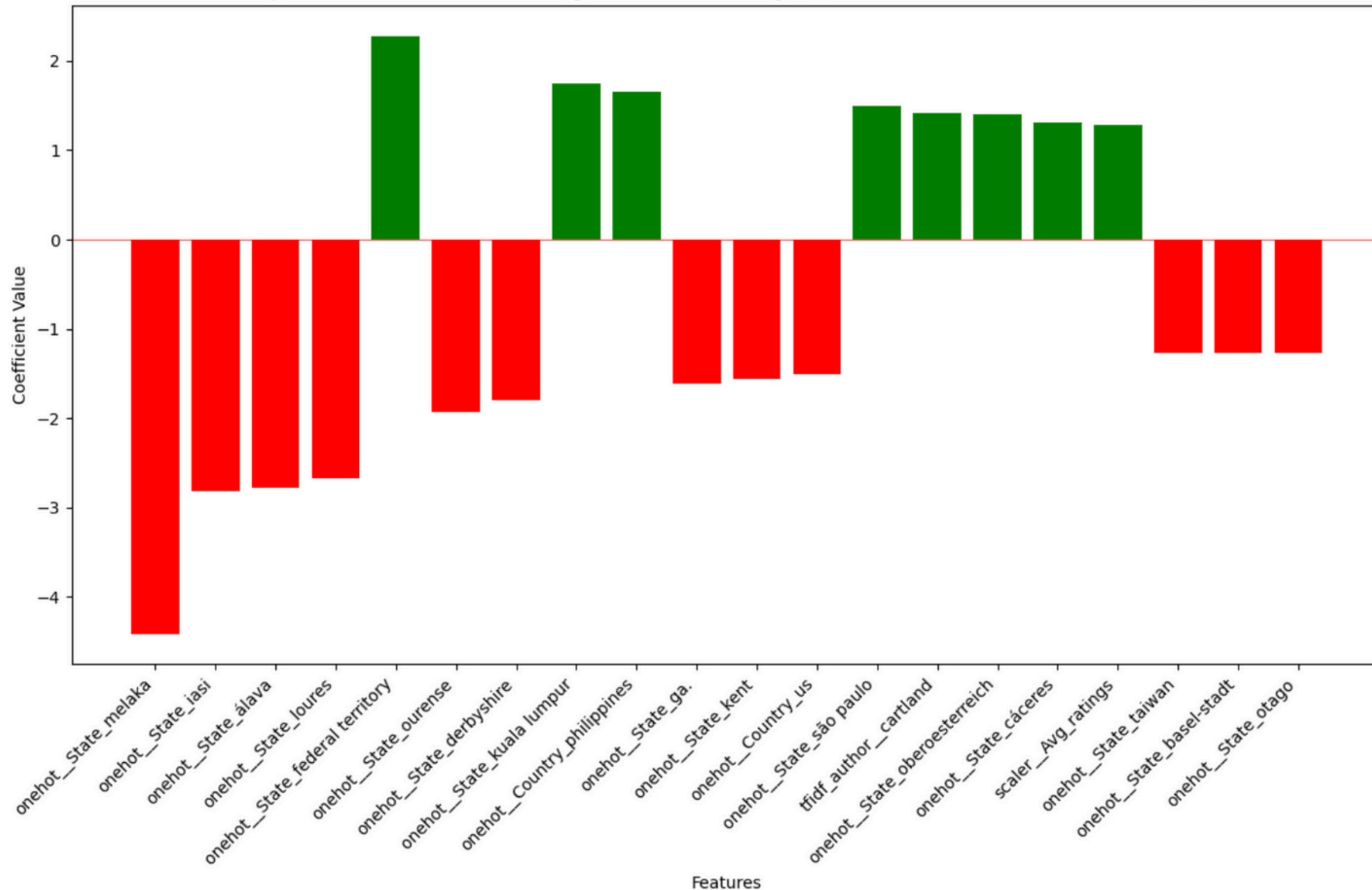
Top 10 Popular Books:

	Title	Author	Publisher	Weighted_Rating
369	Harry Potter and the Prisoner of Azkaban (Book 3)	J. K. Rowling	Scholastic	5.356822
367	Harry Potter and the Goblet of Fire (Book 4)	J. K. Rowling	Scholastic	5.292777
371	Harry Potter and the Sorcerer's Stone (Book 1)	J. K. Rowling	Scholastic	5.087346
368	Harry Potter and the Order of the Phoenix (Book 5)	J. K. Rowling	Scholastic	5.003655
365	Harry Potter and the Chamber of Secrets (Book 2)	J. K. Rowling	Scholastic	4.892268
1267	Ender's Game (Ender Wiggins Saga (Paperback))	Orson Scott Card	Tor Books	4.805739
615	The Little Prince	Antoine de Saint-ExupÃ©ry	Harcourt	4.776057
2842	Harry Potter and the Sorcerer's Stone (Harry Potter (Paperback))	J. K. Rowling	Arthur A. Levine Books	4.665212
466	The Fellowship of the Ring (The Lord of the Rings, Part 1)	J. R. R. Tolkien	Houghton Mifflin Company	4.602030
4975	The Hobbit : The Enchanting Prelude to The Lord of the Rings	J.R.R. TOLKIEN	Del Rey	4.567724

- **Weighted Rating** method: Combines an item's average rating with its the number of ratings
- Does not any content or user related recommendations
- Very generic and not user centric

# LOGISTIC REGRESSION

Top 20 Features: Positive and Negative Effects on Target (Recommended vs. Not Recommended)



## Text and Numeric Processing

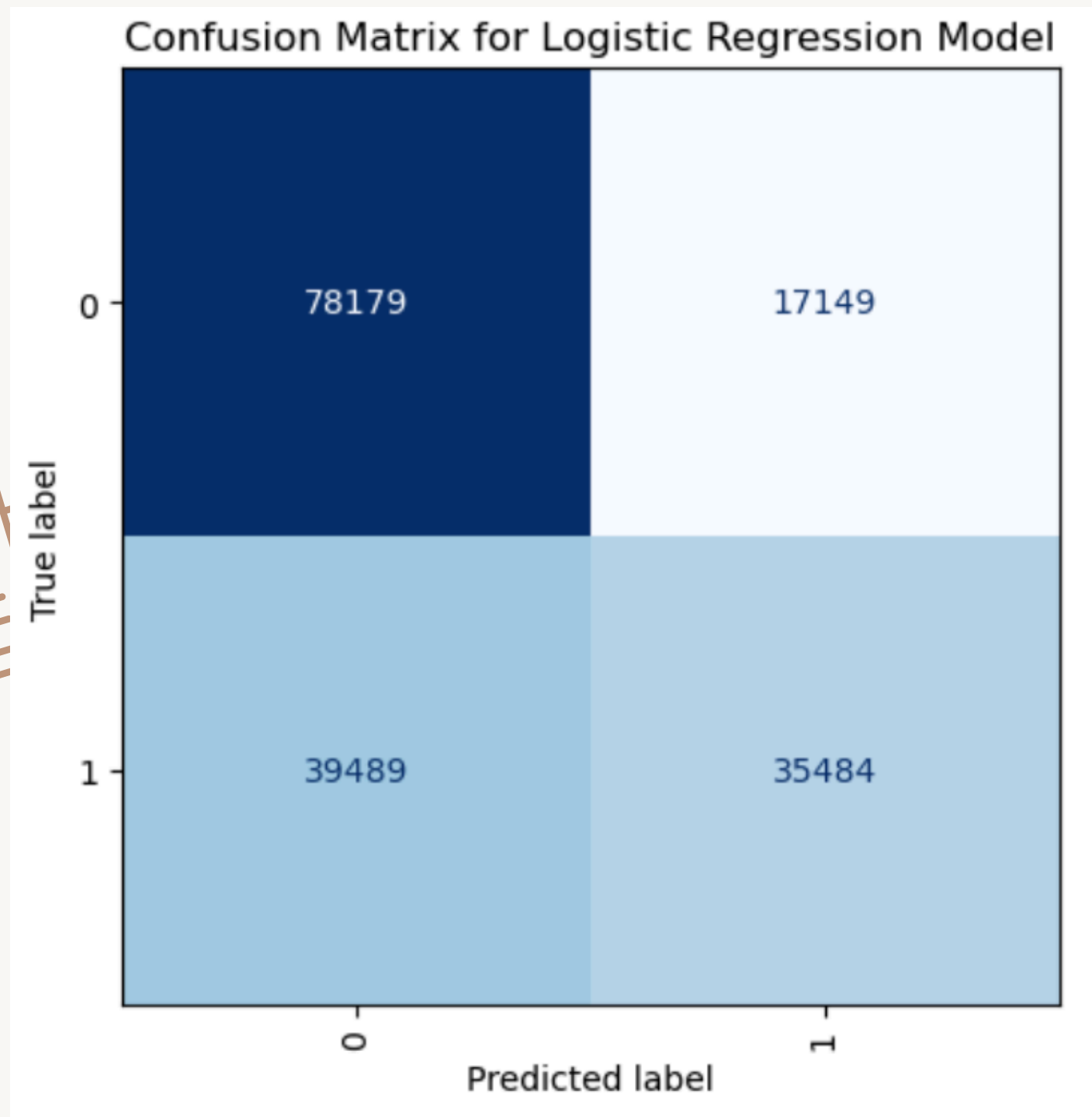
- TF-IDF, one hot encoding
- Standard Scalar

**Train Accuracy: 0.6829**

**Test Accuracy: 0.6674**

- Coefficients show that we have a large effect from States and Countries
- Does not provide much content or user based understanding

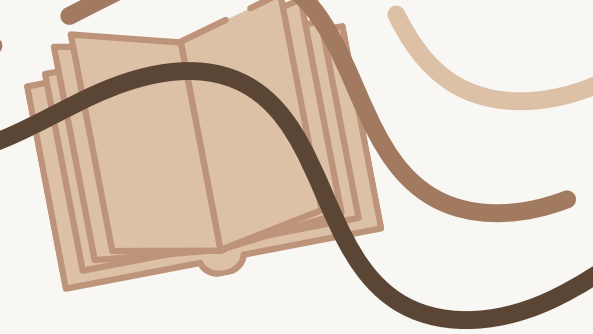
# Evaluating Logistic Regression Model



**Good/Recommend = 1 , Bad/Not Recommend = 0**

- ✓ Accuracy: 75% of predictions are correct
- ⚠ Recall: Only 47% of recommendable cases are identified
- ✓ Precision: 67% of recommendations are accurate
- ⚠ Error Rate: 25% of predictions are incorrect

# Word Embedding with Cosine Similarity



Top 10 similar to 'Harry Potter':

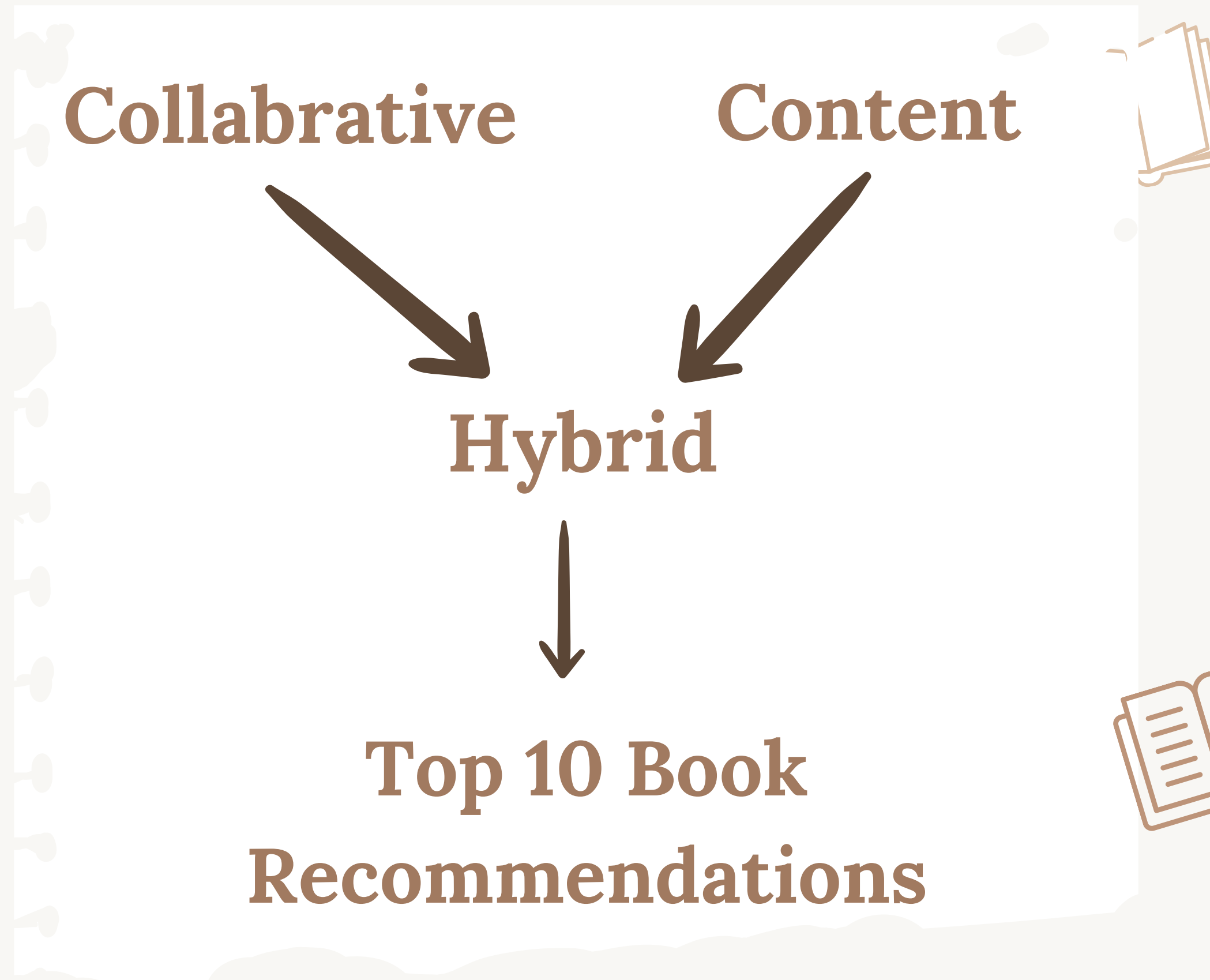
Title	Author	Cosine_Similarity	Avg_ratings
harry potter and the sorcerer's stone	j. k. rowling	0.930866	10.000000
blackeyes	dennis potter	0.928368	3.000000
harry potter and the philosopher's stone	j.k. rowling	0.926405	5.750000
augenstern. roman.	harry mulisch	0.924836	7.000000
harry potter hardcover box set	j. k. rowling	0.923882	5.000000
harry potter and the prisoner of azkaban	j. k. rowling	0.918765	5.850000
harry potter and the goblet of fire	j. k. rowling	0.916542	10.000000
harry potter paperback boxed set	j. k. rowling	0.911517	5.000000
moneymakers.	harry bingham	0.910088	7.000000
harry potter and the chamber of secrets postcard book	j. k. rowling	0.908385	5.680000

- Improved content-based recommendations compared to earlier versions
- Majority of recommended books aligning with user expectations
- Some unrelated books still appear
- A few duplicate editions are still present in the results
- Strong baseline model, especially useful for solving the cold start problem (new users with no history)



# 4 Advance Models

1. **Collaborative Filtering Model**
2. **Hybrid Model** (Weighted/Mixed):
  - Content + Collaborative
  - Adjusts component importance.
  - Presents diverse recommendations





**THANK YOU**