PERSONALIZED BOOK RECOMMENDATION SYSTEM

Course: 15.072 Advanced Analytics Edge

Instructor: Prof. Bart Van Parys

Team Decomposers: Abhranil Chakrabarti, Bibhabasu Das, Shreya Gupta, Rachit Jain

MIT SLOAN SCHOOL OF MANAGEMENT

MIT MANAGEMENT SLOAN SCHOOL

AGENDA

01 INTRODUCTION

02 DATA & EDA

03 RECOMMENDATIONS FOR NEW AND EXISTING USERS

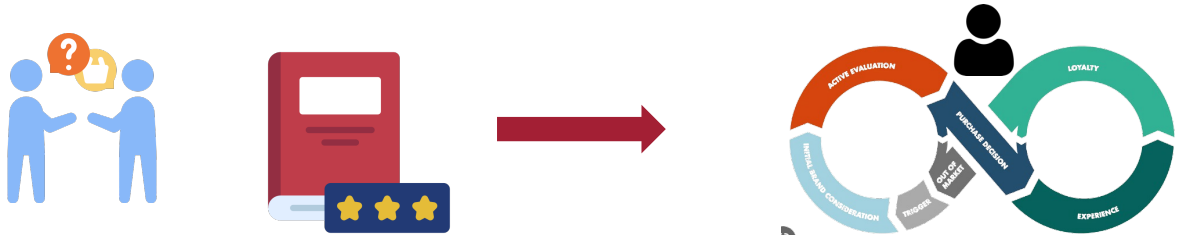
04 RECOMMENDATION VIA OPTIMIZATION

05 RESULTS & IMPACT

MIT MANAGEMENT SLOAN SCHOOL

INTRODUCTION

- Recommendation systems help in understanding user behavior and their requirements to recommend them the offerings/products that are most relevant to them, and hence are most likely to buy.
- The **purpose** of this project is to explore recommendation methods on a specific application, that can be easily extended to any domain, displaying the power and relative ease of implementation of such methods.

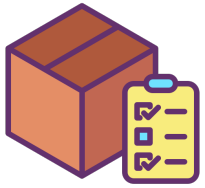


- This project aims to build a **recommendation system** that uses historical book rating information available for users to recommend books to them.
- The key idea behind a recommendation system is that users who share similar preferences are most probable to like similar items; users who like an item are probable to like a similar item.

MIT MANAGEMENT SLOAN SCHOOL

PRIMARY DATA USED

Amazon Data

51.3M reviews **2.9M products**

Sample Review:
ReviewerID, ProductID, Review, Overall Rating, Review Date, etc.
Meta Data:
ProductID, Title, Price, Also_Bought, Also_Viewed, Bought_Together, Category

Kaggle Data



271k books **279k Users** **1.1M Reviews**

Book Data Labels:
ISBN, Book-Title, Book-Author, Publisher, Year-of-Publication, Image-URL
Rating Data Labels:
UserID, ISBN, Rating; User Data: User-ID, Location, Age

MIT MANAGEMENT SLOAN SCHOOL

DATA CLEANING AND PREPROCESSING

FILTERING DATA

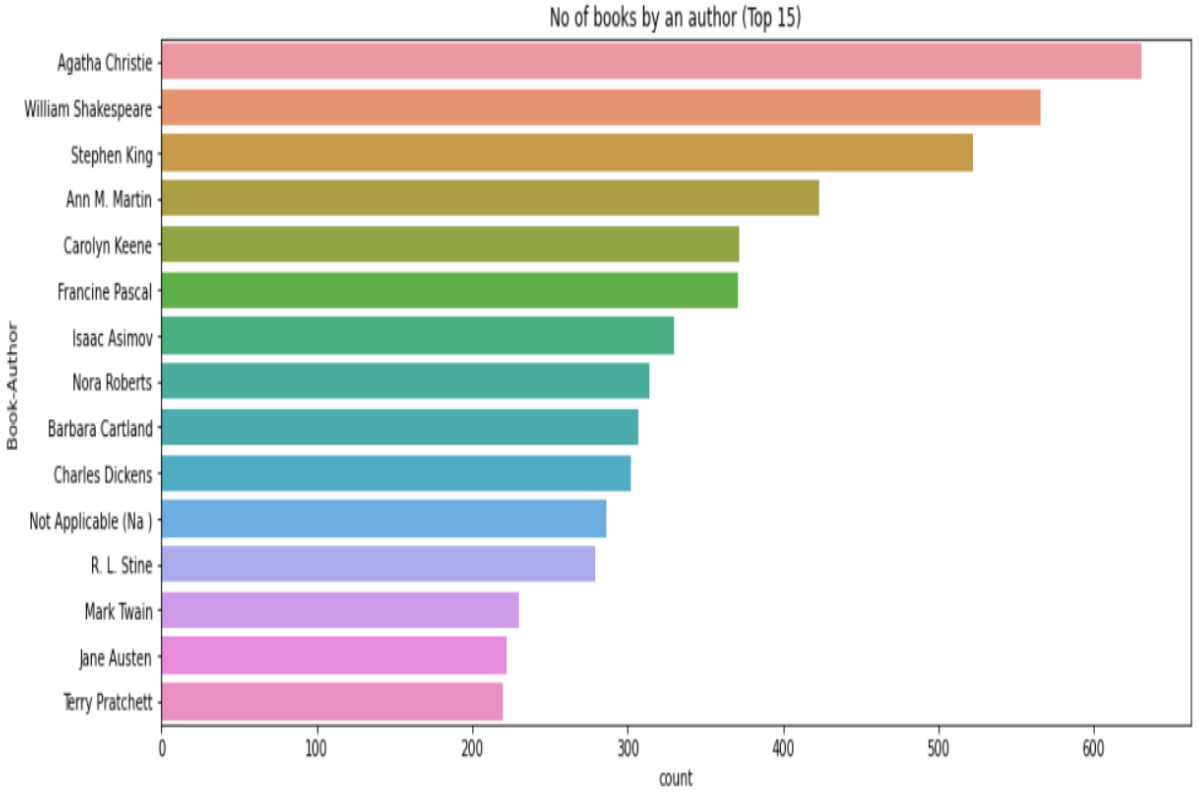
- The user-book rating information is very **sparse,**
- For users who have rated at least 200 books, and
- For books that have at least 50 ratings

EXTRACT BOOK META DATA

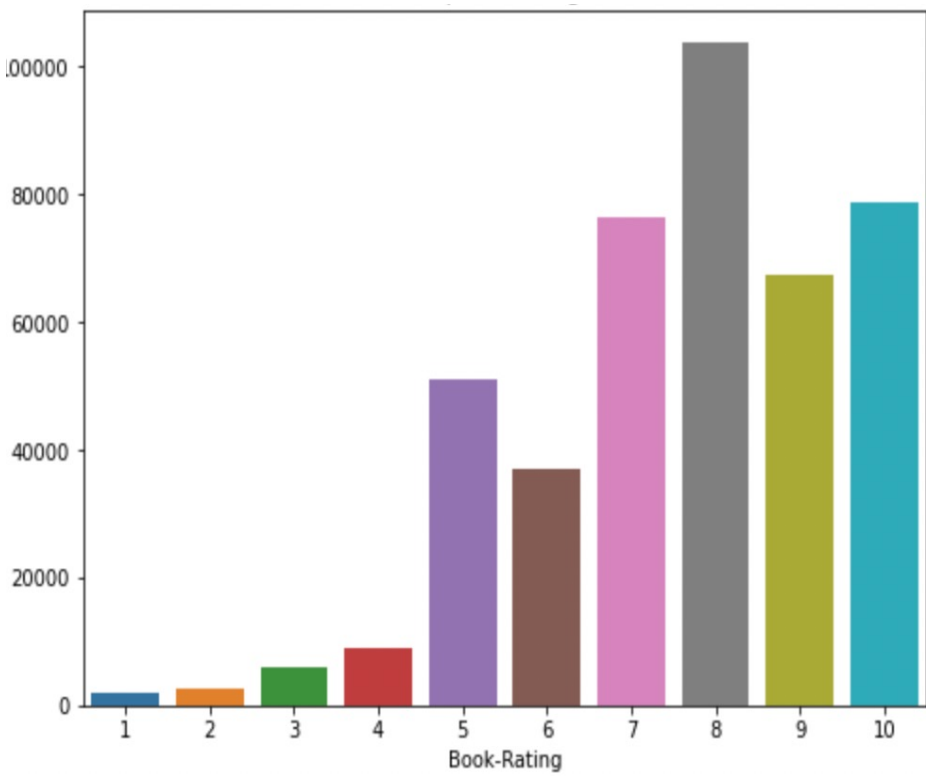
- We extract additional data on books like ‘Description’, ‘Price’ and ‘Category’.
- This was later used for **Text Analytics**

EXPLORATORY DATA ANALYSIS

Number of books by an Author (Top 15)



Number of books rated

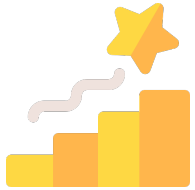


The distribution of ratings indicates that people generally rate a book only if they moderately liked it or really liked it. Very few users bother to rate books if they completely disliked it.

MIT MANAGEMENT SLOAN SCHOOL

RECOMMENDATIONS FOR NEW USERS

POPULARITY BASED RECOMMENDATION



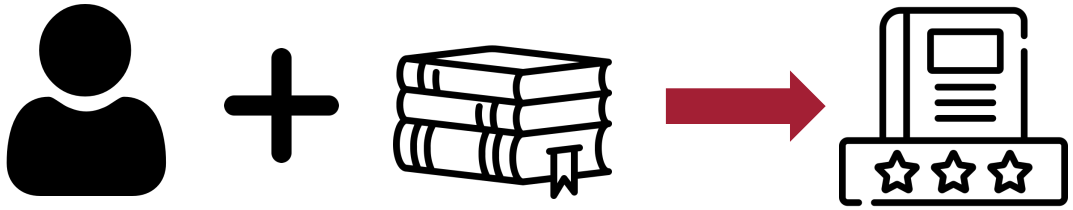
- In a real-time system, the system would recognize this as a new user.
- Start **tracking information** for them to be able to make better and more relevant recommendations in the future.

TEXT ANALYTICS ON BOOK DESCRIPTION

- For a given title, we find the books that are most similar to corresponding books based on how close its description is to other books.
- The evaluation is done in terms of **cosine similarity**.

EXISTING USERS: CONTENT BASED FILTERING

Build models that rely on user and book attributes as inputs to predict ratings. Key techniques used:

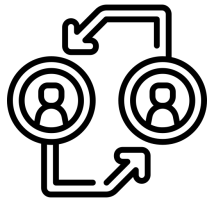


Training Set → Use the Book and user details as model features → Create Regression Model → Obtain rating predictions on test set → Recommend books with highest ratings

Linear Regression	XGBoost	Holistic Regression
MSE TRAINING : 0.909	MSE TRAINING : 0.702	MSE TRAINING : 2.641
MSE TESTING: 0.9133	MSE TESTING: 0.780	MSE TESTING: 2.626

EXISTING USERS: COLLABORATIVE FILTERING

SINGULAR VALUE DECOMPOSITION



We find the users most similar in behavior to current user and recommend the top-rated books by these users

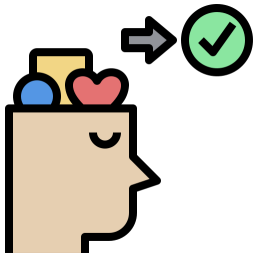
RMSE TESTING: 0.8176

K-NEAREST NEIGHBOUR BASED

Create compressed sparse user-book matrix (for each user and each book) and apply KNN

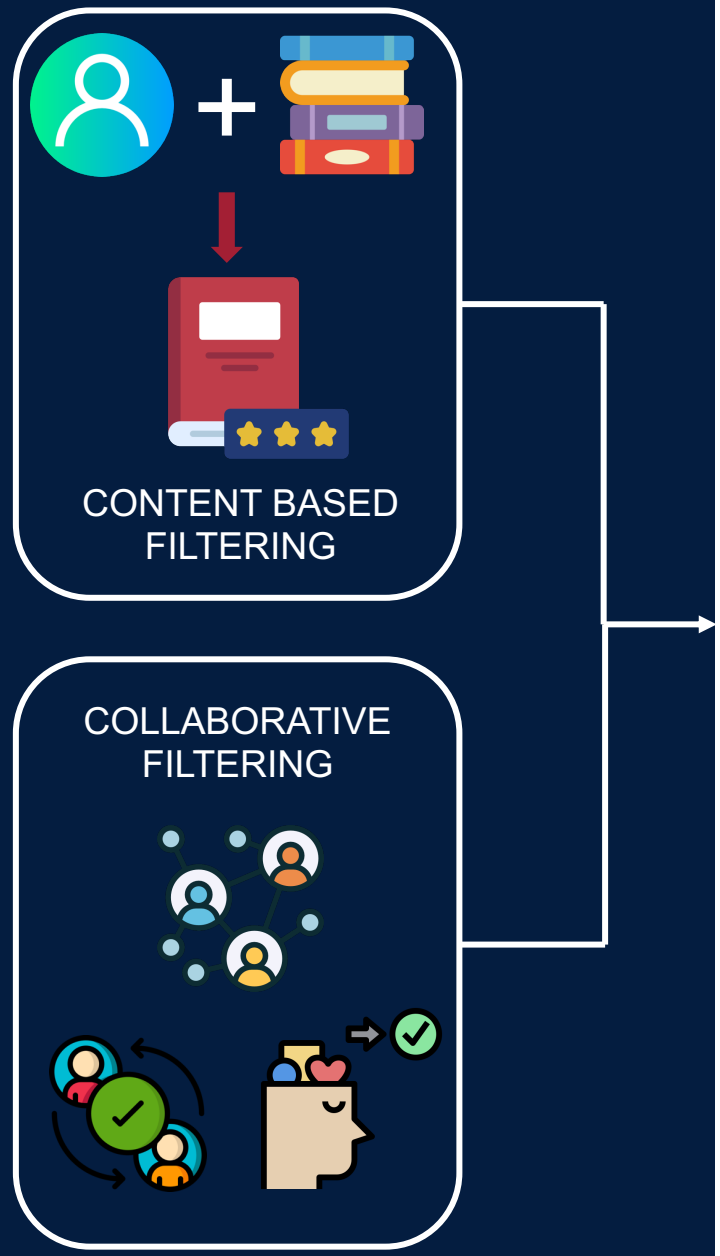
RMSE TESTING: 0.845

ARCHETYPAL USER APPROACH



Classify users into *X* archetypes and based on the archetype into which a new user falls, we recommend the books rated best

RMSE TESTING: 0.721



EXISTING USERS: HYBRID FILTERING

A hybrid recommendation system was built using the combination of both content-based filtering and collaborative filtering systems to recommend the books.

Collective matrix factorization to combine user-item ratings with book and user attributes to give more informed recommendations.

RMSE TESTING : 0.909

MIT MANAGEMENT SLOAN SCHOOL

RECOMMENDATION OPTIMIZATION

For recommendations to a particular user, we solve this optimization problem to account for the liking of the user along with the variety of the recommended book set

We perform the optimization over the 100 most highly rated books of the user to make the problem more tractable

OBJECTIVE FUNCTION

$$\max_X \sum_i X_i * R_i$$

VARIABLES

R_i - Predicted rating of book i by the chosen user
B_i - Feature vector of book i
X_i - Decision Variable: 1 if book i is recommended and 0 otherwise
Y_{ij} - Decision Variable: 1 if book i and book j are both in the recommended

PRIMARY CONSTRAITS

We must recommend K books ($K = 5$ in our case):

$$\sum_i X_i = K$$

Recommended books must not be very similar ($\rho = 0.6$ in our case):

$$Y_{ij} * \frac{B_i^T B_j}{|B_i \parallel B_j|} <= \rho, \forall i, j$$

Linking Constraints:

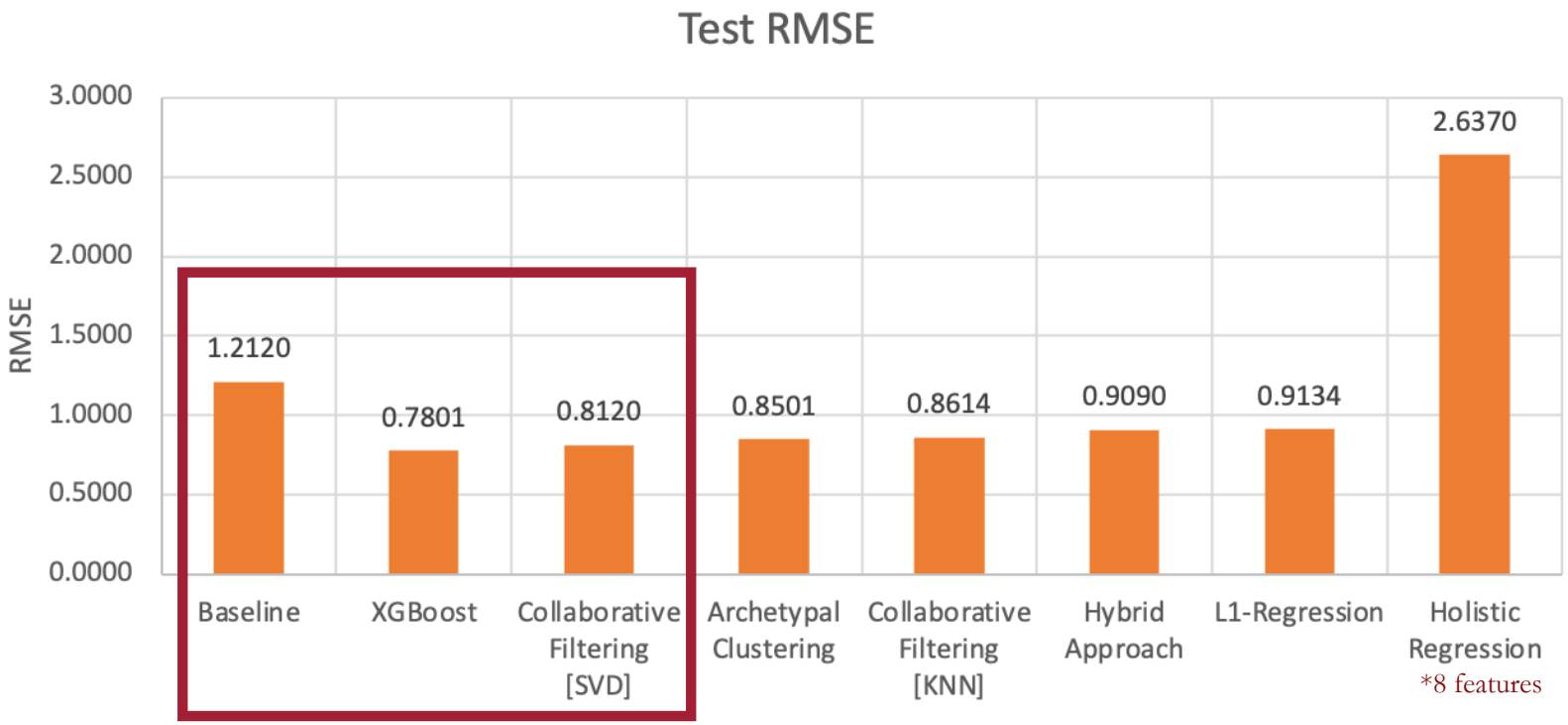
$$Y_{ij} >= X_i, Y_{ij} >= X_j \forall i, j$$

Binary Decision Variables:

$$X_i \in \{0,1\}; Y_{ij} \in \{0,1\}$$

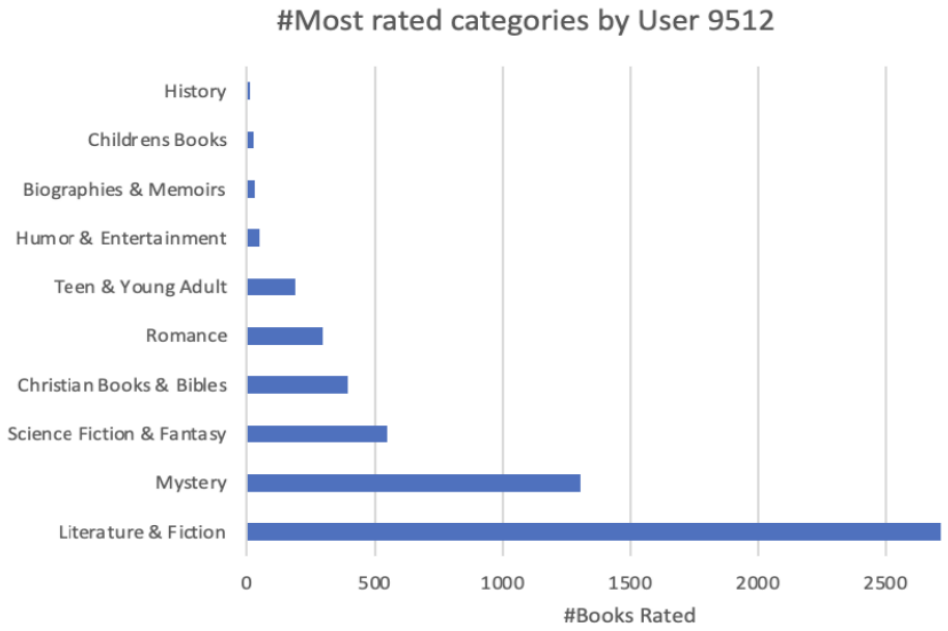
MIT MANAGEMENT SLOAN SCHOOL

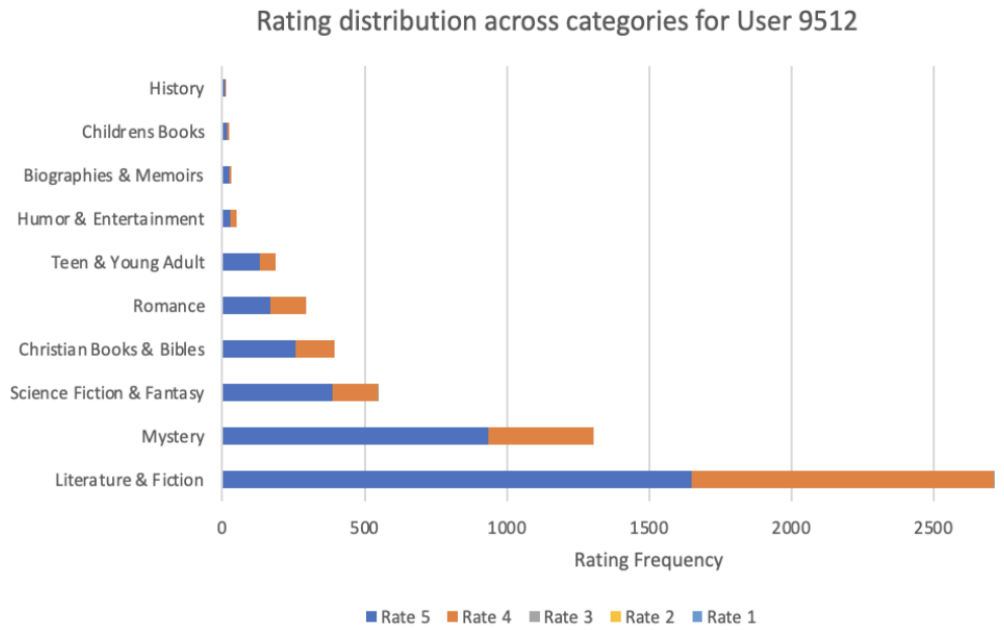
RESULTS



- ✓ **XGBoost** performs the best, followed closely by Collaborative Filtering [SVD] and Archetypal Clustering approach
- ✓ Baseline: Average rating for a user-item pair with added Gaussian noise
- ✓ For Holistic Regression, we only used 8 features to be able to solve the problem in decent time, hence results not that effective.

RECOMMENDATIONS: USER 9512
(MOST NUMBER OF RATINGS)



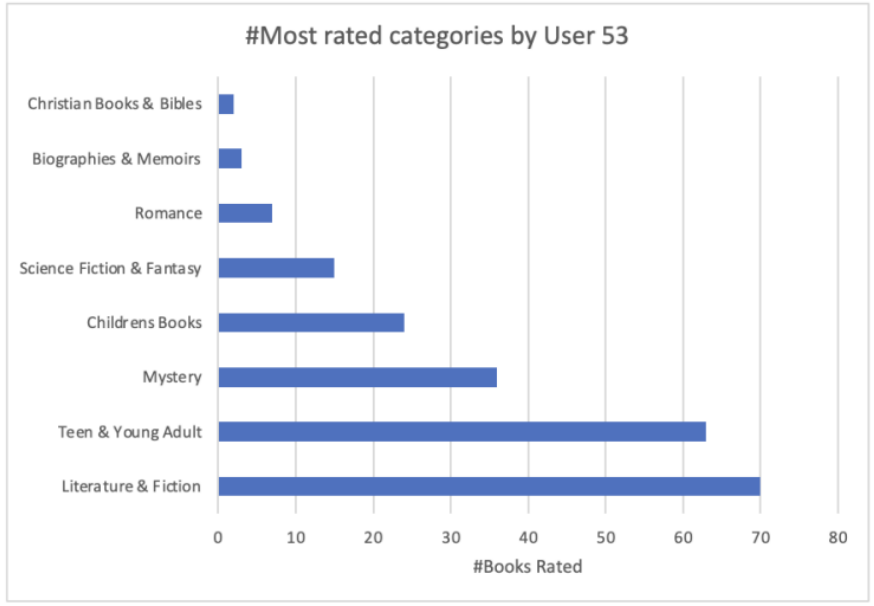


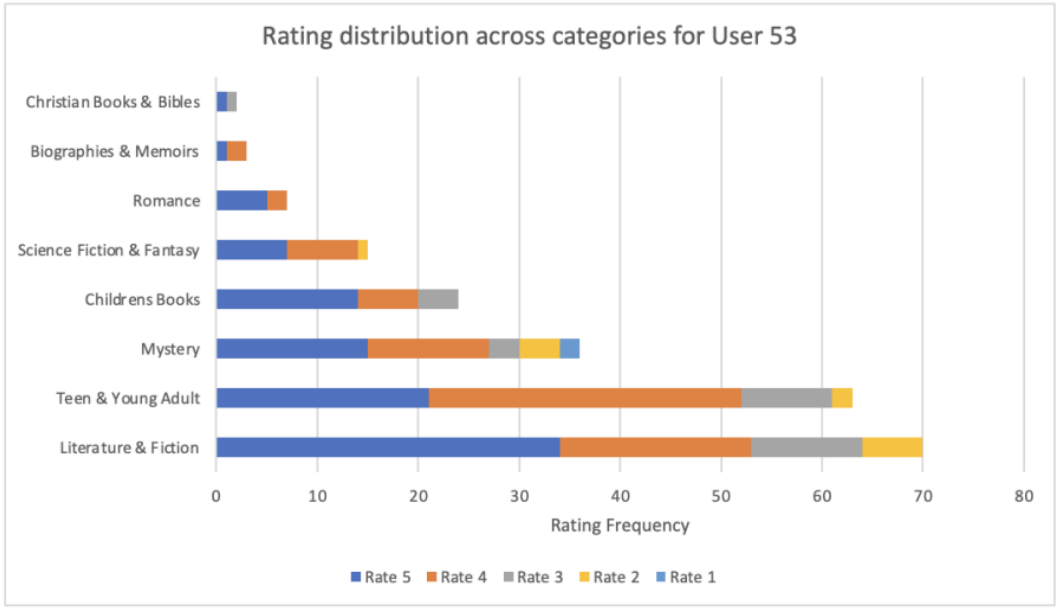
Archetypal Clustering		XGBoost		SVD with Optimization		Hybrid Filtering	
Category	**Title**	**Category**	**Title**	**Category**	**Title**	**Category**	**Title**
Science Fiction & Fantasy	A Dance with Dragons	Mystery	The Fury A Henry Parker Novel	Literature & Fiction	The Winds of War	Travel	A Thousand Days in Venice Ballantine Readers Circle
Literature & Fiction	This Perfect Day	Mystery	Edge of Danger	Biographies & Memoirs	Quartered Safe Out Here	Travel	The Vintage Caper
Politics & Social Sciences	Three Cups of Tea	Mystery	The Bordeaux Betrayal A Wine Country Mystery Wine Country Mysteries	Literature & Fiction	The Long Ships	Travel	A Nail Through the Heart A Novel of Bangkok
Romance	Music of the Heart	Mystery	A Stranger Like You A Novel	Religion & Spirituality	Survival in Auschwitz	Parenting & Relationships	A Boy Should Know How to Tie a Tie And Other Lessons for Succeeding in Life
Crafts	On Paper The Everything of Its TwoThousandYear History	Mystery	Cold Case Alan Gregory	Cookbooks	Dr. Atkins New Diet Revolution	Travel	City of Dark Magic A Novel City of Dark Magic Series

MIT MANAGEMENT SLOAN SCHOOL

RECOMMENDATIONS: USER 53

(RANDOMLY SELECTED)





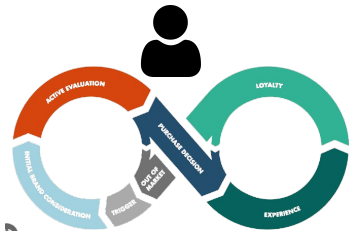
Archetypal Clustering		XGBoost		SVD with Optimization		Hybrid Filtering	
Category	**Title**	**Category**	**Title**	**Category**	**Title**	**Category**	**Title**
Literature & Fiction	Folly	Childrens Books	The Journal of Curious Letters The 13th Reality	Romance	Seduced By Shadows A Novel of the Marked Souls	Childrens Books	Life as We Knew It
Literature & Fiction	Florence of Arabia A Novel	Teen & Young Adult	Incarceron	Romance	Blood on Silk An Awakened By Blood Novel	Childrens Books	The Line
Literature & Fiction	Trespass A Novel	Teen & Young Adult	The Summoning	Romance	Immortal Warrior	Childrens Books	Dead Is a Battlefield
Childrens Books	Click Clack Moo Cows That Type	Teen & Young Adult	The Awakening Darkest Powers	Cookbooks	The GoodtoGo Cookbook	Childrens Books	Trash
Science Fiction & Fantasy	Spin State	Literature & Fiction	The Lucky One	History	The Murder of King Tut The Plot to Kill the Child King	Childrens Books	Dark Secrets 1 Legacy of Lies and Dont Tell

MIT MANAGEMENT SLOAN SCHOOL

IMPACT

EFFICIENT PUSH STRATEGY

This can be used by a marketplace as a part of its push strategy to understand its users, recommend them books while also deciding price point of these books dynamically (based on trending prices in the book category, user willingness to pay).



IMPROVEMENT IN CUSTOMER EXPERIENCE

This system is agnostic to product category and can be embedded as a part of customer experience improvement for any platform, thereby empowering the customer with reliable and personalized information

MIT MANAGEMENT SLOAN SCHOOL

THANK YOU!

MIT
MANAGEMENT
SLOAN SCHOOL