E-commerce Case Study

Maroon Consulting

Jocelyn Chen, Deepali Dagar, John Fitzgerald, Sam Luo, Bruna Medeiros, Kunal Mody

December 9, 2024

Agenda

1. Team Introduction
2. Solution Overview
3. Modeling Process
4. Model Results & Business Value
5. Deployment & Adoption Strategy

Team Introduction

John Fitzgerald
Team Leader

Deepali Dagar
Researcher

Bruna Medeiros
Researcher

Jocelyn Chen
Workflow Coordinator

Sam Luo
Technical Mentor

Kunal Mody
Technical Mentor

About Us

Maroon Consulting

was born out of University of Chicago MS-ADS program in 2014.

MISSION

Our mission is to deliver the best data-driven solutions tailored to each client's unique needs, fostering growth and innovation in their endeavors.

VALUES AND CULTURE

Through an environment of collaboration, meritocracy, and excellence, we aim to empower organizations to make informed decisions and achieve lasting success.

Our Impact and Expertise

McDonald's All-Day Breakfast

- Identified key drivers in a regional promotion and optimized marketing strategies to boost ROI.
- Coordinated with CMO and Regional Leads to roll-out marketing plan.

Fitness Center Retention

- Improved member retention and minimized churn rates through data-driven strategies.
- Minimized churn by 20% compared to baseline in the year after implementation.

Verizon Impact Application

- Accurately forecasted customer default probabilities, reducing financial losses and maximizing gains.
- Developed a predictive, data-driven application to enhance economic impact improving original model's ROI by $27 million per 1 million applicants.

Background

Background

The book e-commerce market is experiencing remarkable growth, projected to reach USD 23.12 billion by 2026, with a compound annual growth rate of 4.9%[1]. Digital transformation and changing consumer behaviors are reshaping how books are discovered, purchased, and consumed.

Main Challenge

Despite the availability of extensive customer data, many online bookstores still rely on basic **"most popular" recommendation systems**, with **many readers reporting frustration** from receiving suggestions for books they've already purchased.

Business Opportunity

- Enhance customer engagement
- Increase average order value
- Create meaningful connections with readers
- Drive personalized marketing campaigns

1: AppleWorld.Today. (2021, February 10). *Global e-book market expected to see strong growth through 2026*. AppleWorld.Today. https://appleworld.today/2021/02/global-e-book-market-expected-to-see-strong-growth-through-2026/

Problem Statement

Problem Statement

Background

Current recommendation system treats all readers identically, missing opportunities to leverage rich customer data for better recommendations and increased sales.

Understanding Our Customers

- **What** they prefer (e.g., genres or product categories).
- **How often** they shop, **what** they buy, and **how much** they spend.

Model Evaluation and Promotion

Success will be measured through increased conversion rates, higher average order values, and improved customer engagement metrics across segments.

Opportunities

Higher Sales

By focusing on the **top 15%** of engaged customers, we can achieve significant **revenue growth** (estimated **+20%**) while also driving small but meaningful improvements across other segments

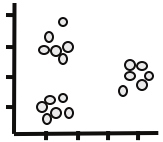
Solution Summary

Segmentation

Data Processing & Cleaning

Identified data **patterns**/**trends** like most **popular genres**

Detected **outliers** and **monetary skewness**

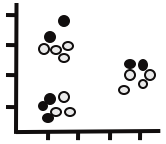


First K-means Clustering

Clustering based on:

1. **Recency:** Time since last purchase.
2. **Frequency:** Number of purchases.
3. **Monetary:** Total money spent.

Resulted in **4** total clusters

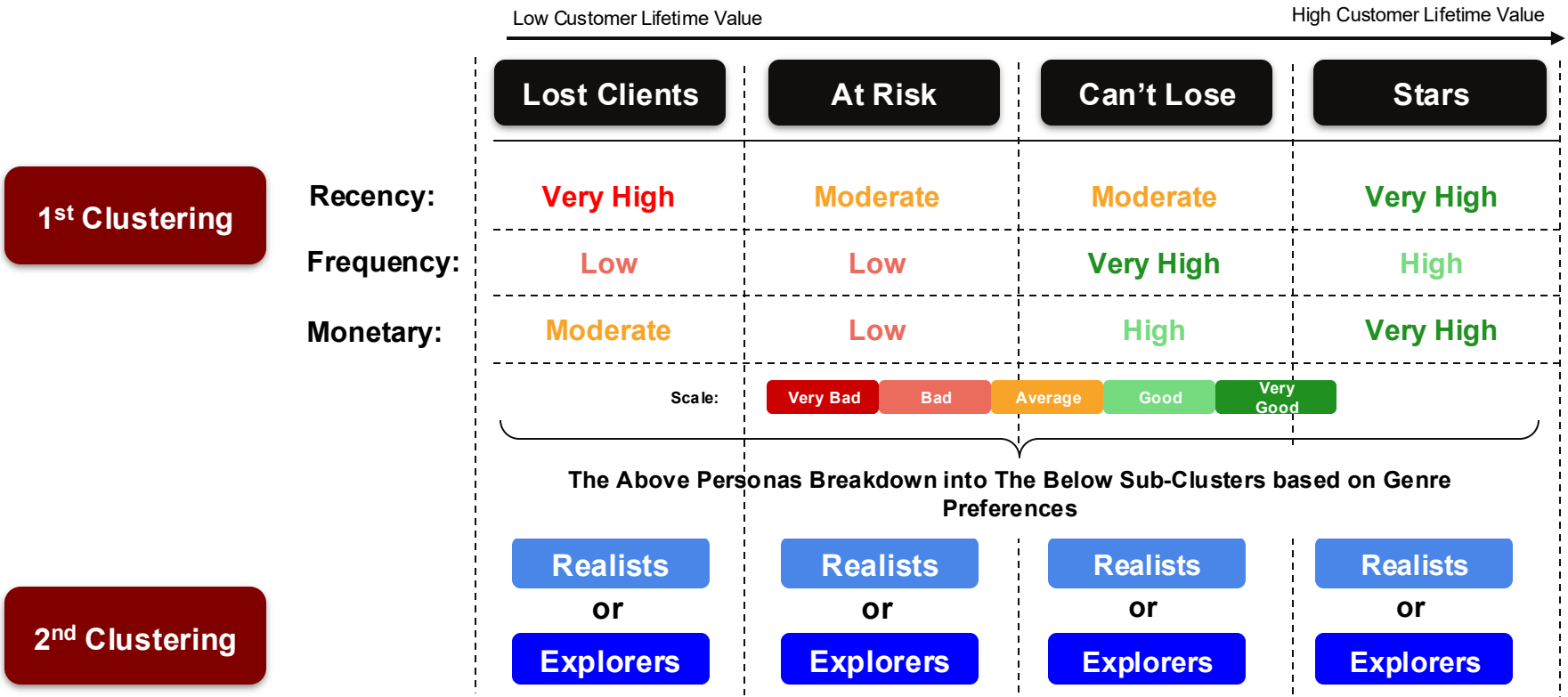


Second K-means Clustering

Clustering based on **each client's preferred genre.**

Creates 2 clusters for **each** initial cluster (total of **8** clusters)

Personas


Low Customer Lifetime Value
High Customer Lifetime Value
Lost Clients
At Risk
Can't Lose
Stars
1st Clustering
Recency:
Very High
Moderate
Moderate
Very High
Frequency:
Low
Low
Very High
High
Monetary:
Moderate
Low
High
Very High
Scale:
Very Bad
Bad
Average
Good
Very Good
The Above Personas Breakdown into The Below Sub-Clusters based on Genre Preferences
2nd Clustering
Realists
or
Explorers
Realists
or
Explorers
Realists
or
Explorers
Realists
or
Explorers

1st Level Clustering

The first level of analysis used to create “personas” captures the behavior that characterizes the Customer Lifetime Value (CLV) of each customer

Low Customer Lifetime Value → High Customer Lifetime Value

	Lost Clients	At Risk	Can’t Lose	Stars
Recency:	Very Low	Moderate	Moderate	Very High
Frequency:	Low	Low	Very High	High
Monetary:	Moderate	Low	High	Very High

Scale: Very Bad | Bad | Average | Good | Very Good

2nd Level Clustering

The second level of analysis used to create “personas” captures the unique interests of customers and is used for enhance personalization beyond their CLV characteristics.

Realists

Clients interested in **every-day**, **mundane**, **common** genres:

Examples:

- Sport
- Game Riddles
- Economy
- Cartoons
- Politics
- Non-Books
- Health
- Science

Explorers

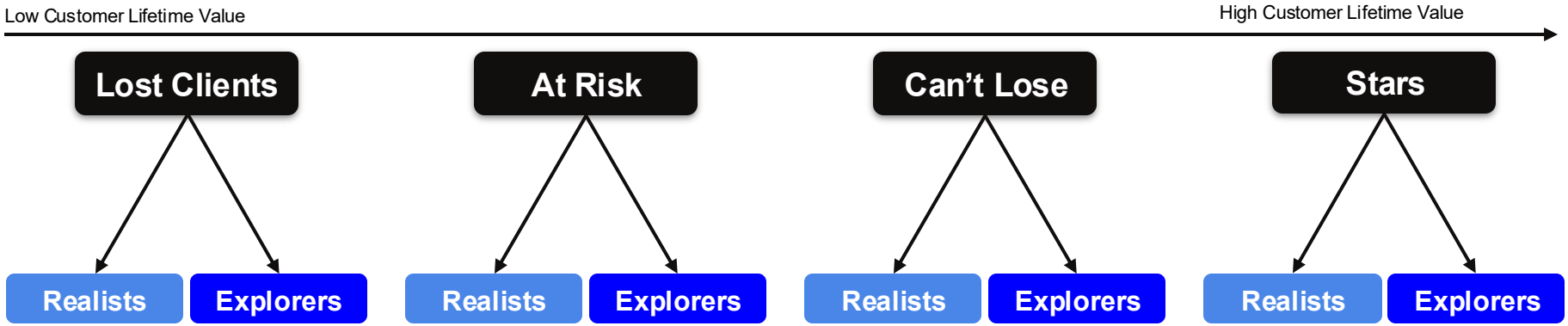
Clients interested **humanities** or **travel** related genres:

Examples:

- Learning
- History
- Travel Guides
- Maps
- Contemporary History
- Religion
- Encyclopedia
- Linguistics

Summary of Personas

Eight clusters in total: **1st Level Clustering:** 4 clusters
2nd Level Clustering: 2 Sub-Clusters, each



Strategic Marketing:

We will recommend a specific **discount** and **message** for each of the four clusters.
Then, for each sub-cluster, adjust the **recommended genres** to their preference.

Combined Marketing Messaging

Strategic Marketing:

We will recommend a specific **discount** and **message** for each of the four clusters.
Then, for each sub-cluster, adjust the **recommended genres** to their preference.

	Lost Clients	At Risk	Can't Lose	Stars
Discount	**1-5% off**	**2% off or personalized offer to re-engage**	**3% off or Loyalty Rewards (points of free shipping)**	**Early access** to releases of **exclusive bundles**
Message	"We miss you! Reignite your curiosity with captivating reads in [Preferred Genres]!"	"Don't miss out! Rediscover your favorite reads and enjoy personalized offers tailored for you."	"Thank you for your loyalty! Enjoy these exclusive recommendations, tailored just for you, to keep your reading journey exciting!"	"As one of our top readers, you deserve the best! Check out these exclusive offers curated just for you!"
Suggestion	**Preferred genres and/or new genres.**	**Related genres.**	**Related genres.**	**Premium books or rare editions of favorite genres.**
Rationale	Customers that have not engaged in a long time. They need an attractive discount and a personalized push to **remind them of the value they can derive from your bookstore.**	Showing signs of churn with moderate recency and low frequency. A personalized offer with related genres can **prevent them from leaving entirely**.	Were highly active and frequent buyers. Losing them would have a **disproportionate impact on sales**. Therefore, recommend new, but related genres of preference to **keep them interested**.	Customers who spends significantly and are highly engaged. They **value exclusivity and premium experiences**, so ensuring those are maintained is essential.

Business Value

Optimized Strategy Increases Revenue by 4.29%

Top 15%

Current Revenue	Estimated Revenue	Revenue After Discount
$1,278,100	$1,533,725	$1,533,724

Remaining customers

Current Revenue	Estimated Revenue	Revenue After Discount
$7,214,021	$7,507,883	$7,324,240

Current Revenue: **$8,493,129**

Predicted Revenue: **$8,857,964**

Revenue: $364,836*

*Specific calculations and assumptions can be found in the appendix

Implementation Strategy

A **Pilot Deployment**

- Collaborate with CMO and Webmaster to create pilot deployment strategy.
- Start with **top 15% most engaged customers.**
- Tailor recommendations based on the customer segment.

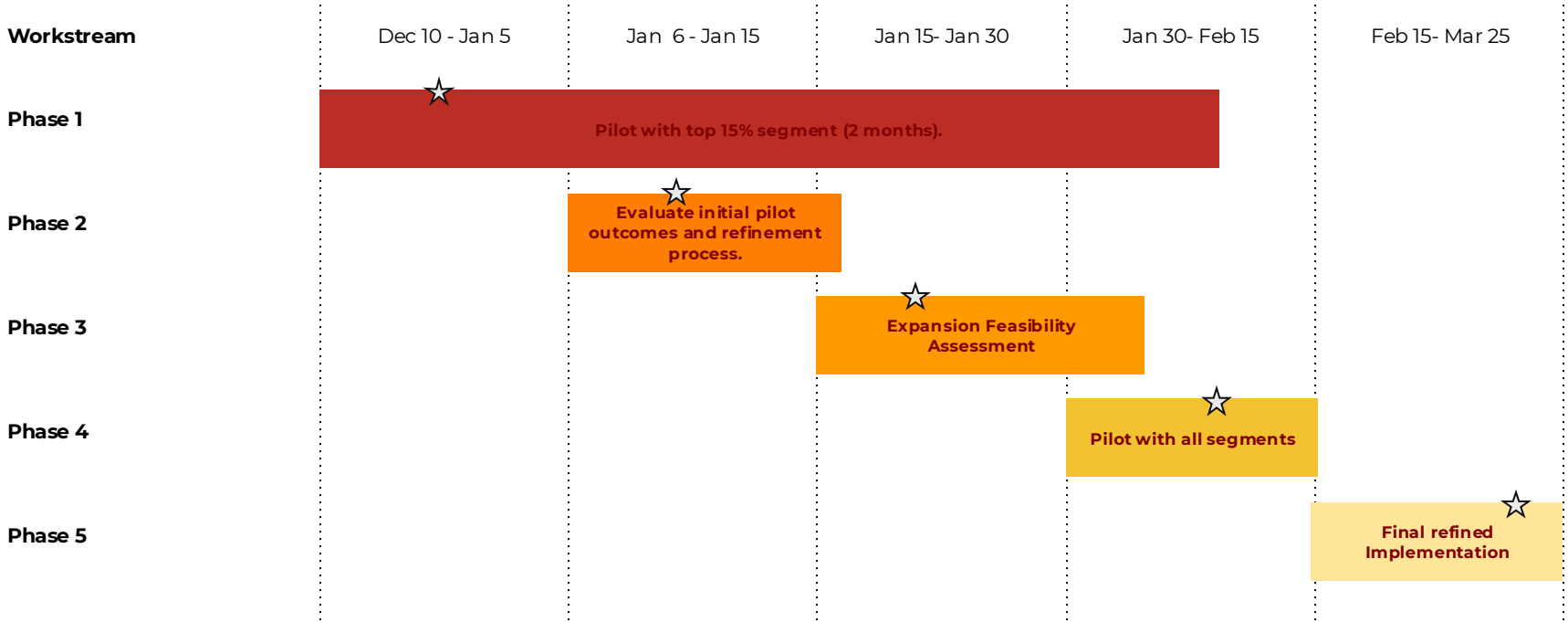
B **Refine the Recommendation Engine**

- Collect performance data from the pilot.
- Measure KPIs such as **conversion rate uplift, AOV, and retention**.
- Phased rollout with **A/B testing.**
- Continuous monitoring and optimization.

C **Data Integration**

- Automate data pipelines to **update clusters dynamically as new data arrives**.
- Ensure scalability and minimal downtime during deployment.
- Set up automated monitoring and reporting systems

Timeline



☆ Key Check-in Meetings

Dec 20: Ensure readiness for the pilot launch, addressing key metrics, and initial segmentation setup.
Jan 10: Assess early pilot results and refine strategies based on customer engagement insights.
Jan 20: Evaluate final pilot results to assess feasibility for broader rollout..
Feb 5: Discuss scalability of the approach and plan for full implementation.
Mar 20: Finalize rollout strategies and gain full approval for complete implementation.

Summary

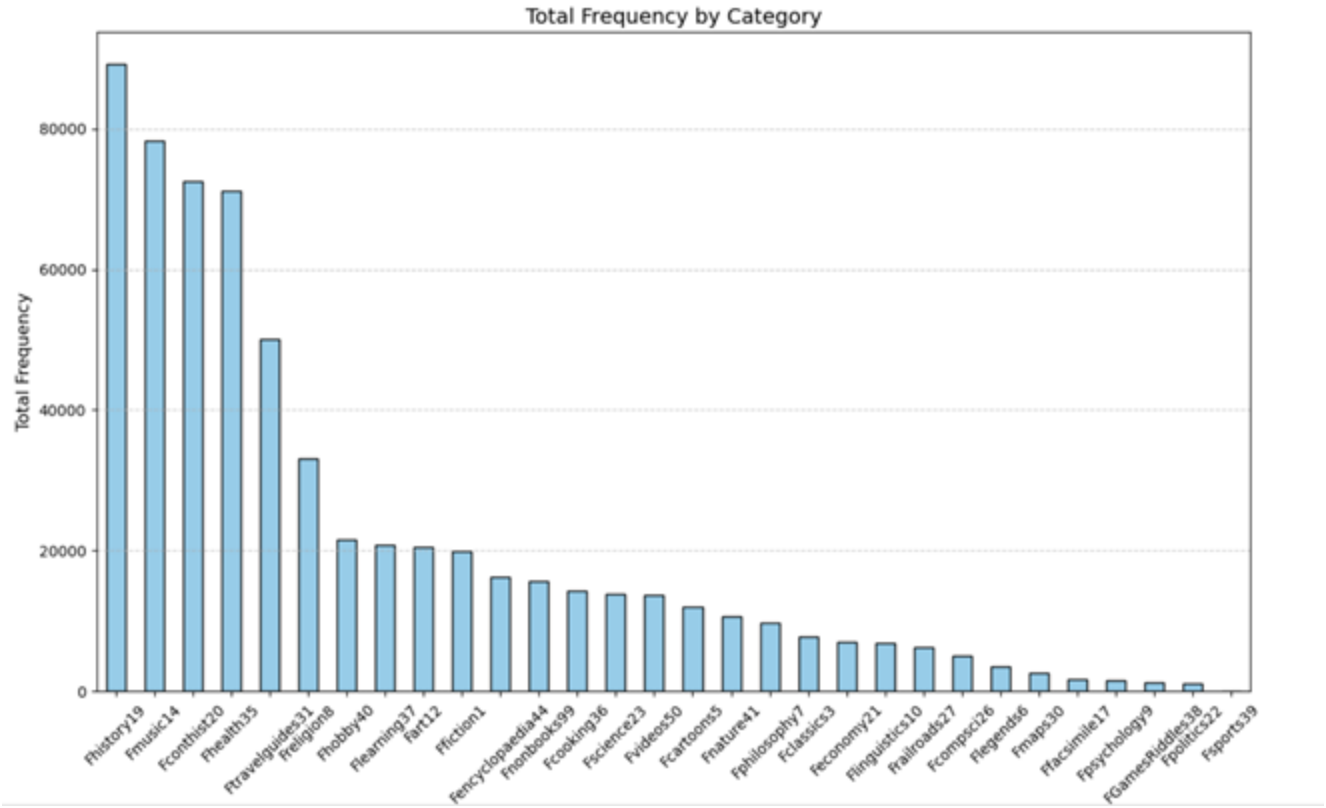
Our Solution	• Building an RFM model based on K-means for classifying. • Identify the consumption intentions of different customer groups and tailor sales strategies to suit local conditions.

Business Value	• The estimated additional revenue is $ 364,835 which brings significant economic benefits. • Personalized recommendation algorithms help to enhance user retention and the perception of user consumption.

Implementation	• Conduct an algorithm pilot among the top 15% customers. • Establish a timeline for optimizing and improving the algorithm, and gradually roll out the algorithm to more consumers.

Appendix

Distribution of Genres



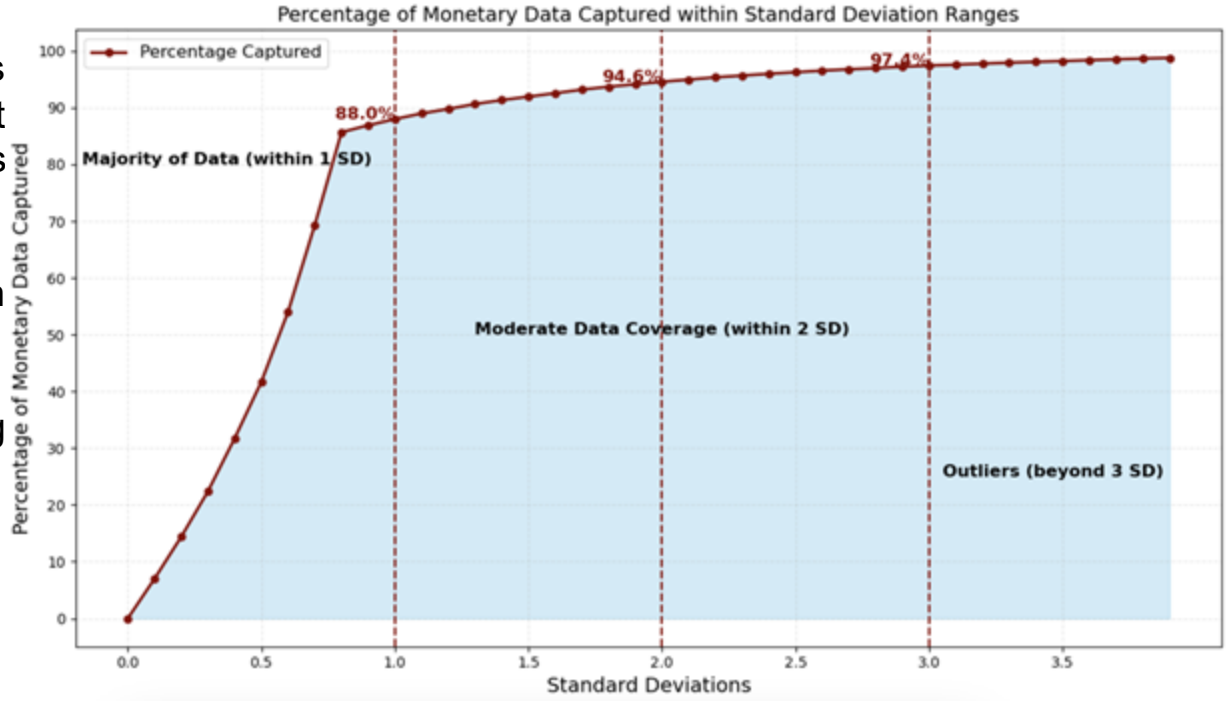
Top Sellers: History, Music, Contemporary history dominate sales and revenue.

Balanced Demand: Genres like hooby and learning show consistent performance.

Low Performers: Niche genres (e.g., Maps, Sports) contribute minimally, offering growth potential.

Outlier Analysis

- Distribution of customer sales (monetary values) is hugely right skewed, with a majority of values concentrated towards lower sales
- Outliers may have an impact on clustering models
- Elbow method for determining cutoff for outliers
- Most data (88%) captured lies < 1 standard deviation away



EDA/Insights

Monetary (variable “m”)

Range: $0 to $532,892

Breakdown of 'm' (count)

Bins	Percentage
<500	89.5%
500-1000	7.5%
1000-2000	2.4%
2000-3000	0.4%
3000-4000	0.1%
4000-5000	0.0%
>5000	0.1%

Breakdown of '<500' Bin

Bins	Percentage
<100	47.7%
100-200	20.9%
200-300	10.4%
300-400	6.4%
400-500	14.5%

Insights:

1. **Significant Spending Concentration:** 89.5% of clients have spent less than $500 in total, with nearly half (47.7%) spending under $100, indicating a heavily skewed distribution toward lower-value customers.

1. **Customer Segmentation Pattern:**
   - **Core Low-Value Segment:** 47.7% spend under $100
   - **Mid-Value Segment:** 31.3% spend between $100-300
   - **Premium Segment:** Only 0.6% spend over $3000

EDA/Insights

Frequency (variable “f”)

Range: 0 to 118

Breakdown of 'f' (Frequency)

Bins	Percentage
<20	95.9%
20-40	3.6%
40-60	0.4%
60-80	0.1%
80-100	0.1%
>100	0.1%

Breakdown of '<20' Bin

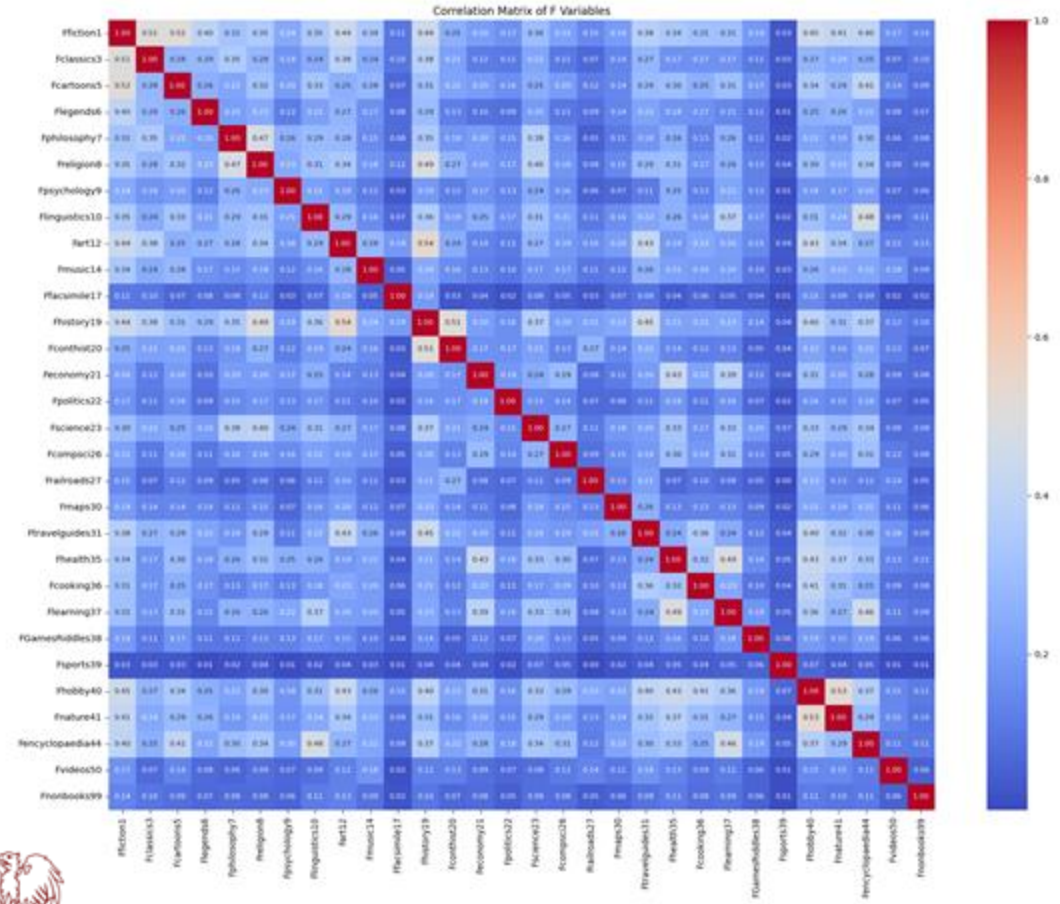
Bins	Percentage
0-5	47.7%
5-10	20.9%
10-15	10.4%
15-20	6.4%

Insights:

1. **Highly Infrequent Usage Pattern:** 95.9% of clients engage less than 20 times, showing a heavily concentrated pattern of low-frequency usage.

1. **User Engagement Breakdown::**
   - **Majority Low-Frequency Users:** 47.7% engage only 0-5 times
   - **Occasional Users:** 20.9% engage 5-10 times
   - **Regular Users:** 10.4% engage 10-15 times
   - **Highly Active Users:** Only 0.2% engage more than 80 times

EDA/Insights: Correlation Matrix for f Variables

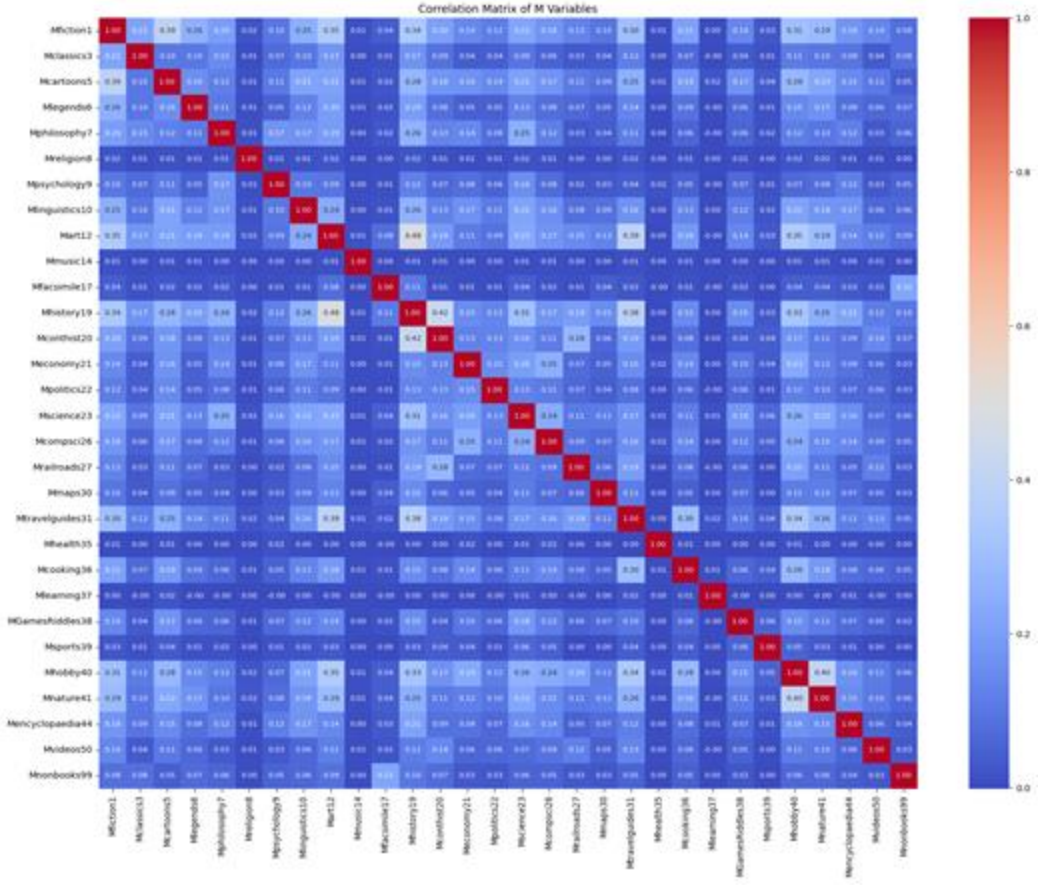


The lighter the color, the higher the correlation.

It can be observed that variables 1, 10, 19, and 44 have a higher correlation with other variables.

These correlations suggest that some customers may purchase other items when buying a particular product, and if we can grasp these relationships, it will undoubtedly generate additional commercial benefits.

EDA/INSIGHTS Correlation Matrix for M Variables



The lighter the color, the higher the correlation. It can be observed that variables 18 and 40 ave a higher correlation with other variables.

The correlation matrix of F and M shows differences, indicating that they each contain independent information, which will have separate effects during clustering.

Model Selection

Selecting **K-Means** for creating a recommendation system based on customer segments is the preferred option. With a silhouette score of 0.50, there are moderate cluster performance, but K-Means is highly interpretable for the business to create personalized recommendations and is MECE, meaning that all new customers will be assigned to a group.

Model	MECE (Y/N)	Strengths	Weakness
DBSCAN	No	Detects arbitrary shapes and outliers	Struggles with high-dimensional, sparse e-commerce data; not ideal for large datasets
GMM	No	Handles overlapping clusters with probabilistic assignments	Less interpretable for business stakeholders; slower with large datasets.
Light GBM	Yes	Classification	Not designed for unsupervised segmentation; requires labeled data for training
K-Means	Yes	Simple, efficient, and interpretable	Assumes spherical clusters; may oversimplify customer behavior in complex datasets

Building the Best Clustering for Business Outcomes

	Step	Description	Outcome
1.	High Level RFM Clustering	Utilized K-Means on Recency, Frequency, Monetary Value, and Time On File being open	Enables strategic allocation of marketing based on customer lifetime value derived from RFM
2.	Sub-Clustering Based on Genre Preferences	Within each High Level RFM Cluster, clustered on Genre Categorizes	Provides deeper insights into customer tastes and interests for more refined targeting
3.	Identify Unique Personas	Used the differentiators within Sub-Clusters to find unique attributes for customers in them	Highlights unique behaviors and attributes that can be used by business units
4.	Create Personalized Marketing Strategy & Recommendations	Identified the unique products, messaging, and offers to recommend the personas	More effectively design tailored marketing, product recommendations, etc.

Persona Results

Step 1: Clustering based on **R, F**, **M** and **Time on File** metrics, resulting in 4 clusters.
Step 2: Divide each into 2 subclusters each, based on preferred **genres**, totaling 8 clusters.

	GROUP	LOST CLIENTS		CAN'T LOSE	
Step 1:	Recency	Very High		Moderate	
	Frequency	Low		Very High	
	Monetary	Moderate		High	
	SUB-GROUP	Explorers	Realists	Explorers	Realists
Step 2:	Genre Preference	1. History 2. Health 3. Music 4. Travel Guide 5. Contemporary History	1. Sports 2. Politics 3. Facsimile 4. Games Riddles 5. Psychology	1. History 2. Health 3. Travel Guides 4. Contemporary History 5. Religion	1. Non-books 2. Sports 3. Games Riddles 4. Politics 5. Psychology
	Discount	10% off		Loyalty rewards (points or free shipping), or discounts	
	Message	"We miss you! Reignite your curiosity with captivating reads in [Preferred Genres]!"		"Thank you for your loyalty! Enjoy these exclusive recommendations, tailored just for you, to keep your reading journey exciting!"	
	Suggestion	Preferred Genres.		Related Genres: Art, Linguistics, Learning and Maps.	Related Genres: Cartoons, Economy.
	Rationale	Customers that have not engaged in a long time. They need an attractive discount and a personalized push to remind them of the value they can derive from your bookstore.		Were highly active and frequent buyers. Losing them would have a disproportionate impact on sales. Therefore, recommend new, but related genres of preference to keep them interested.	

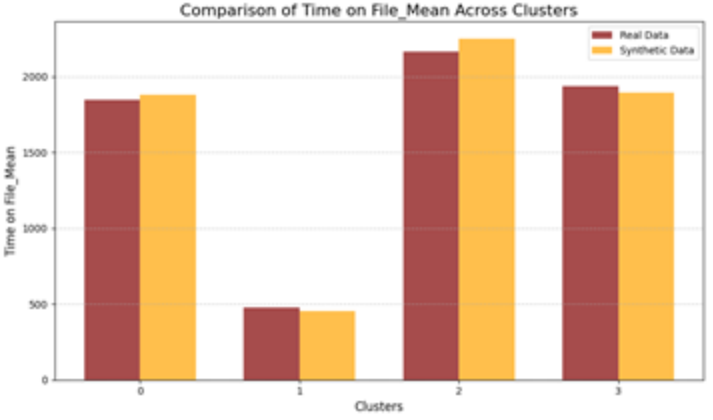
Persona Results

Step 1: Clustering based on **R, F**, **M** and **Time on File** metrics, resulting in 4 clusters.
Step 2: Divide each into 2 subclusters each, based on preferred **genres**, totaling 8 clusters.

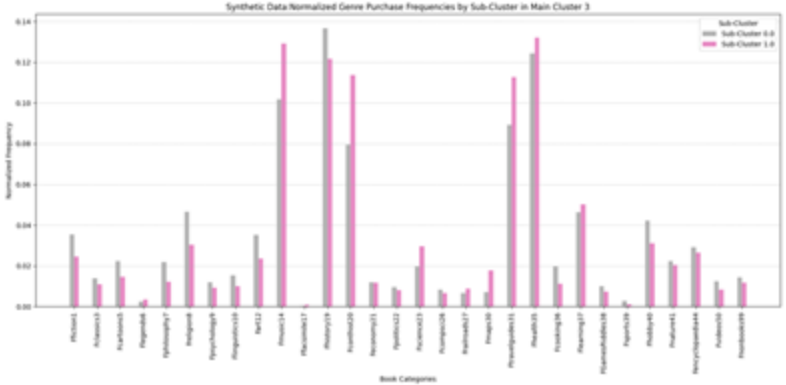
	GROUP	STAR CUSTOMERS/CHAMPIONS		AT RISK	
Step 1:	Recency	Very Low		Moderate	
	Frequency	High		Low	
	Monetary	Very High (outlier)		Low	
	SUB-GROUP	Realists	Explorers	Explorers	Realists
Step 2:	Genre Preference	1. Sports 2. Facsimile 3. Politics 4. Maps 5. Games Riddles	1. Health 2. Music 3. History 4. Contemporary History 5. Travel Guides	1. Learning 2. Travel Guides 3. Music 4. Encyclopedia 5. Religion	1. Health 2. Economy 3. Videos 4. Psychology 5. Non-Books
	Discount	Early access to new releases or exclusive bundles, but no discounts (focus on exclusivity).		5% off or a personalized offer to re-engage.	
	Message	"As one of our top readers, you deserve the best! Check out these exclusive offers curated just for you!"		"Don't miss out! Rediscover your favorite reads and enjoy personalized offers tailored for you."	
	Suggestion	Premium books or rare editions in their favorite categories.		Related Genres: Art, Linguistics, and Philosophy.	Related Genres: Science, Politics, Learning.
	Rationale	Customers who spends significantly and are highly engaged. They value exclusivity and premium experiences, so ensuring those are maintained is essential.		Showing signs of churn with moderate recency and low frequency. A personalized offer with related genres can prevent them from leaving entirely.	

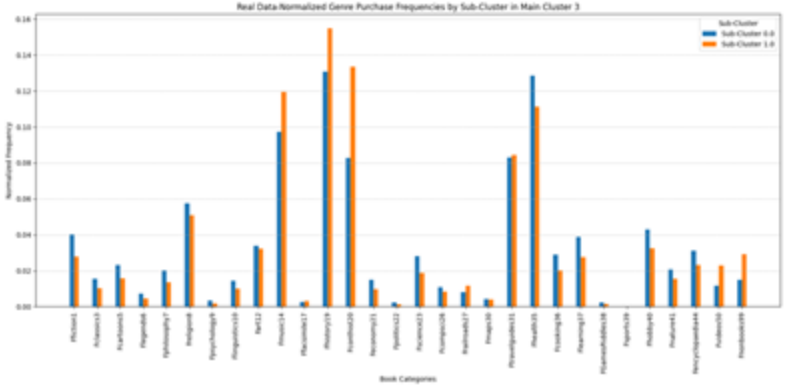
Next Steps - Synthetic Data Validation

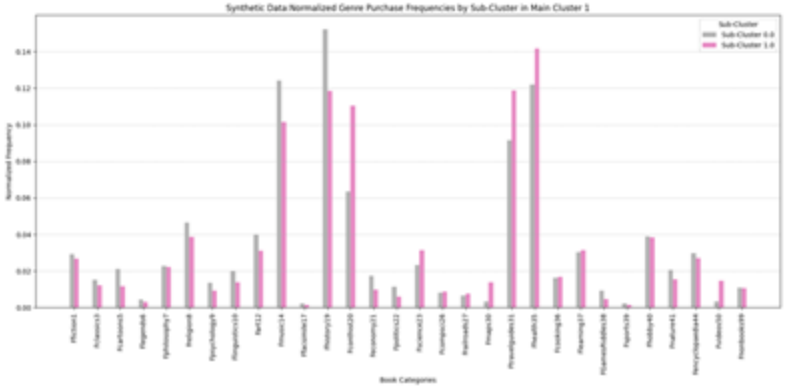
1. Purpose:
    a. Validate model performance using realistic synthetic data that mimics real-world distributions.
2. Method:
    a. Technique:Conditional Tabular GAN
    b. Process:
        i. Trained a generative model to learn data distributions.
        ii. Generated realistic synthetic data for validation.
3. Result:
    a. Clusters in synthetic data align closely with those in real data.
    b. The result demonstrates robustness of the model across diverse datasets.

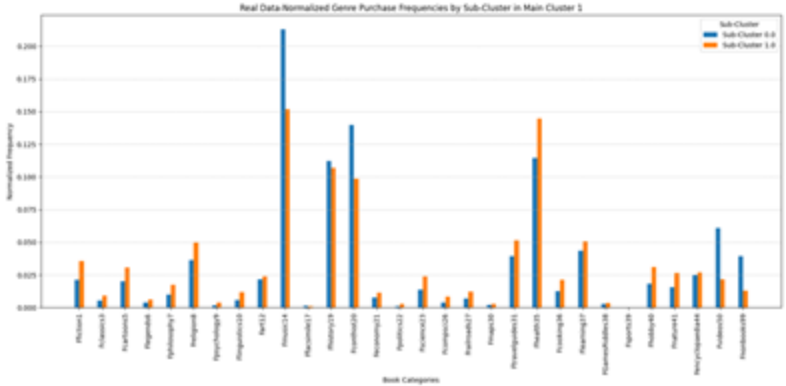


Synthetic Data vs Real Data

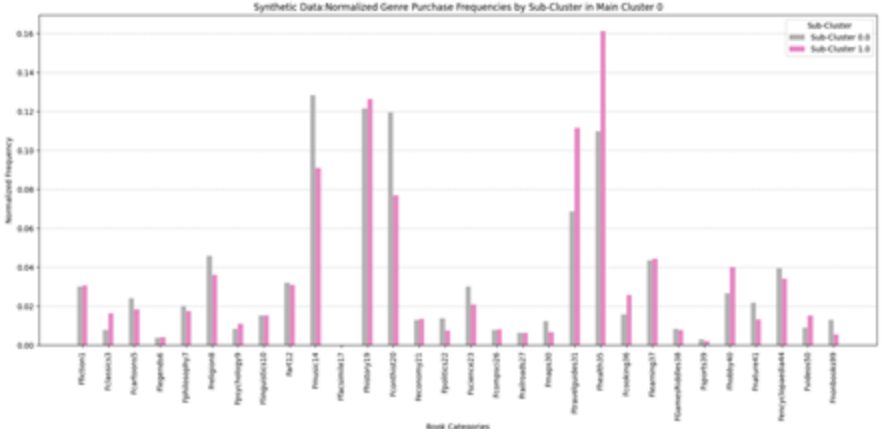


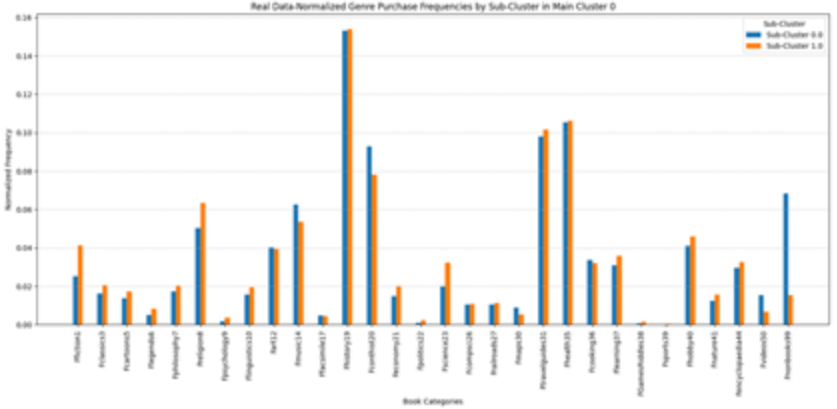


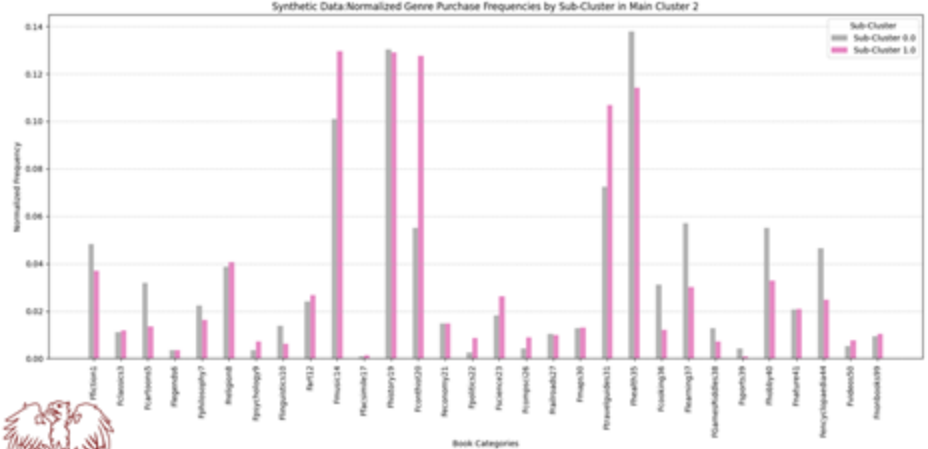


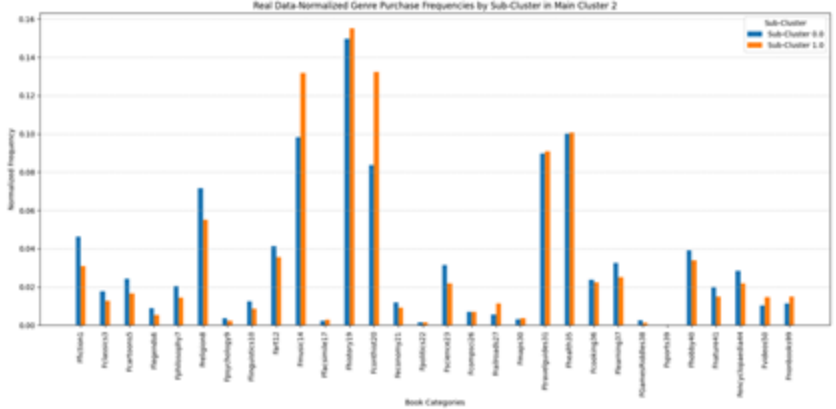


Synthetic Data vs Real Data









Synthetic data- Sample

index	id	r	f	m	tof	Ffiction1	Fclassics3	Fcartoons5	Flegends6	Fphilosophy7	Freligion8	Fpsychology9	Flinguistics10	Fart12	Fmusic14	Ffacsimile17	Fhistory19	Fconthist20	Feconomy21	KMeans_4_Cluster
0	4463512	1066	10	1504.71	2286	2	1	1	1	1	6	0	1	3	0	2	2	0	1	**3**
1	1282610	7	2	41.21	42	4	2	1	1	1	4	0	1	1	0	2	7	0	1	**1**
2	645137	1189	3	222.09	2139	2	1	1	1	1	1	0	1	2	0	0	2	0	1	**0**
3	977195	227	17	345.91	2010	5	2	1	2	1	2	0	1	2	0	1	9	0	1	**2**
4	16093261	363	6	134.65	2005	2	1	1	1	1	1	0	1	1	0	2	10	0	2	**3**
5	2332064	1387	6	714.63	2381	2	1	2	1	1	2	0	1	1	1	2	5	0	1	**0**

Revenue per persona

Group	SUBGROUP	Current Revenue	% of total revenue
Stars	Explorers	$331,438.06	3.9%
	Realists	$632,206.89	7.4%
Can't lose	Explorers	$1,424,224.59	16.8%
	Realists	$3,487,611.27	41.0%
At Risk	Explorers	$636,092.51	7.5%
	Realists	$532,891.85	6.3%
Lost Clients	Explorers	$947,145.14	11.1%
	Realists	$501,514.76	5.9%
	TOTAL	$8,493,124.07	100%

We expect to deliver +20% incremental sales among the most engaged 15% of customers and 2-5% among other customers.

Business Value

Top 15%

Group	SUBGROUP	Current Revenue	% Increase*	Estimated revenue	Discount	Revenue After discount
Stars	Explorers	$331,438.06	20%	$397,725.67	0%	$397,725.67
	Realists	$632,206.89	20%	$758,648.27	0%	$758,648.27
Can't lose	Explorers	$91,220.56	20%	$109,464.67	0%	$109,464.67
	Realists	$223,238.13	20%	$267,885.75	0%	$267,885.75
	TOTAL	$1,278,103.64		$1,533,724.36		$1,533,724.36

Predicted revenue: $1,533,724.36

–

Current revenue: $1,278,103.64

⇩

△ revenue: $255,620.72

* Note that 'Can't Lose Customers' are categorized into two groups: those in the top 15% and those in the remaining 75%.

Business Value

Remaining customers

Group	SUBGROUP	Current Revenue	% Increase*	Estimated revenue	Discount	Revenue After discount
Can't lose	Explorers	$1,333,004.03	5%	$1399654.23	3%	$1357664.60
	Realists	$3,263,373.14	5%	$3426541.79	3%	$3323745.53
At Risk	Explorers	$636,092.51	3%	$655175.28	2%	$642071.77
	Realists	$532,891.85	3%	$548878.60	2%	$537901.02
Lost Clients	Explorers	$947,145.14	2%	$966088.04	1%	$956427.15
	Realists	$501,514.76	2%	$511545.05	1%	$506429.59
	TOTAL	$7,214,021.43		$7,507,882.99		$7,324,239.66

Predicted revenue: $7,324,239.66

−

Current revenue: $7214021.43

⇩

△ revenue: $110218.23

* Note that 'Can't Lose Customers' are categorized into two groups: those in the top 15% and those in the remaining 75%.

Business Value

Top 15%

Current Revenue	Estimated revenue	Revenue After discount
$1,278,103.64	$1,533,724.36	$1,533,724.36

Remaining customers

Current Revenue	Estimated revenue	Revenue After discount
$7,214,021.43	$7,507,882.99	$7,324,239.66

Predicted Revenue:
$8,857,964

Current Revenue:
$8,493,129

Revenue: $364,835

Optimized Strategy Increases Revenue by 4.29%