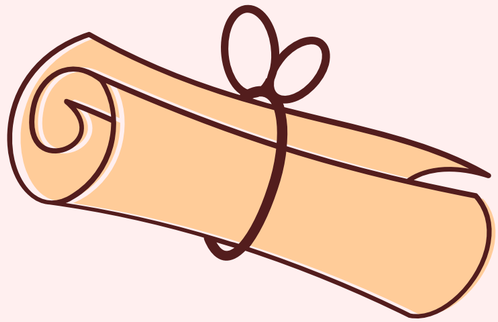
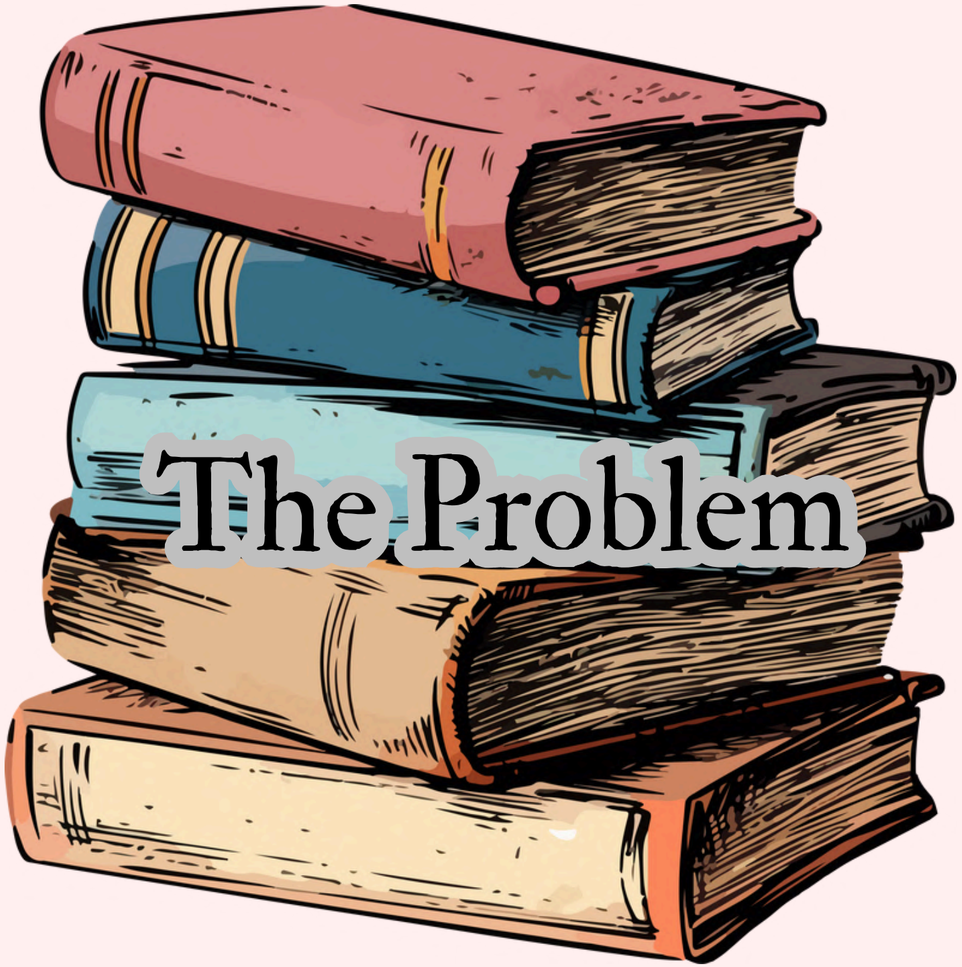
DATA-DRIVEN LIBRARY ACQUISITION SYSTEM WITH INTEGRATED SHELF OPTIMIZATION

By- Akshat Mishra, Mukul Jangra, Sher Partap Singh





The Problem

The Problem: Modern Library Challenges

📚 Rapid growth of library collections is outpacing available shelf space

🧱 Physical space constraints in university and institutional libraries

📉 Traditional acquisition methods rely on:

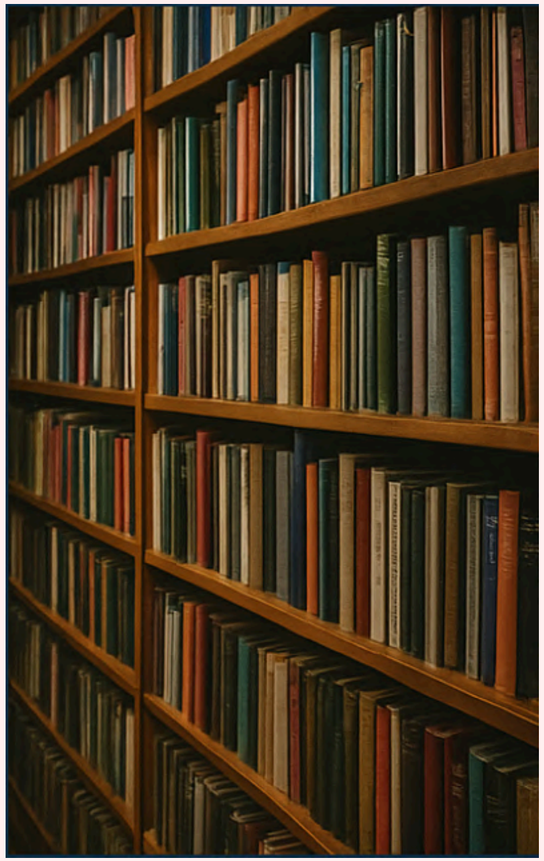
- Librarian expertise
- Publisher-driven models
- Basic circulation statistics

🔁 Frequent reshuffling of collections wastes time and resources

❄️ Cold-start problem: No data to assess interest in new books

🎓 Student behavior is evolving:

- Borrowing less
- Downloading from the internet
- Sharing materials among peers



Problem Statement:

Libraries today are challenged by growing collections, space constraints, and shifting user behavior. Traditional acquisition and shelf strategies often result in underused resources.

There is a critical need for a data-driven framework that:

- Predicts user interest for smart acquisition
- Optimizes shelf placement based on predicted usage
- Effectively handles cold-start issues using contextual features

This integrated approach provides a unified solution for acquisition and shelf planning—improving user satisfaction and optimizing resource utilization.

The Gap:

- Fragmented Solutions:

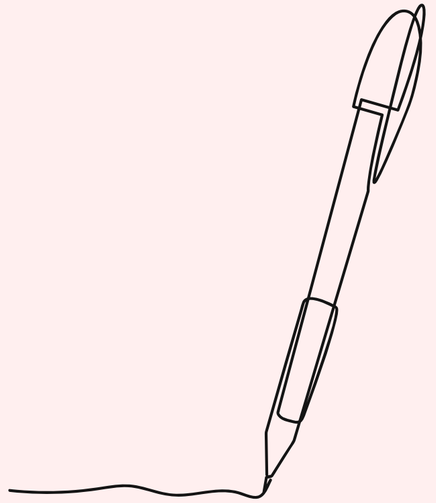
Existing models focus independently on either acquisition (DDA, EBA) or recommendation, not both.

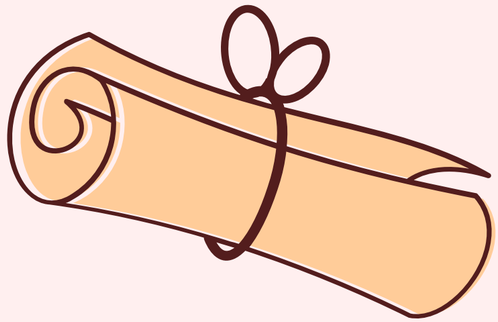
- Lack of Integration:

Most systems do not combine user preference prediction with physical shelf optimization in a single, adaptive framework.

- No Use of Latent & Contextual Features Together:

Collaborative filtering captures latent user patterns, while modern ML (like XGBoost) allows rich feature input — yet no system merges them effectively for library management.





Literature Review

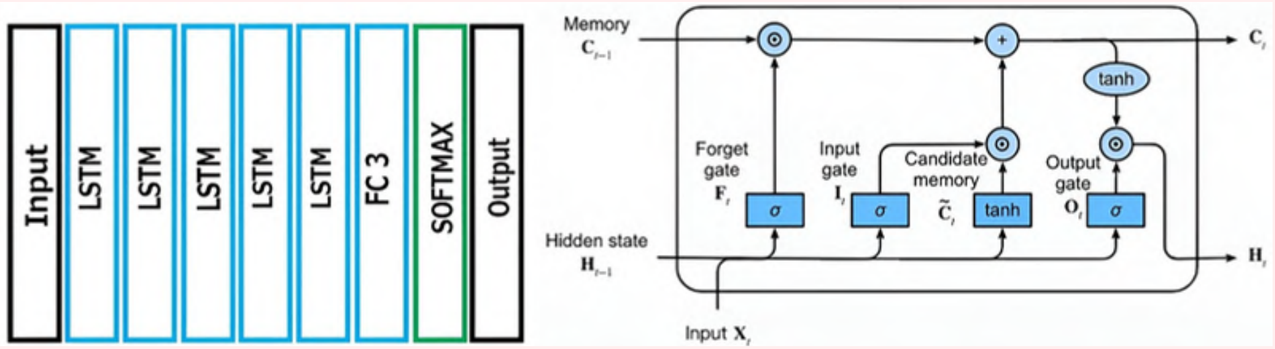


Figure : Recommendation system with LSTM architecture.

Paper 1

Personalised Book Recommendation Systems Using Collaborative Filtering

Performance Metrics

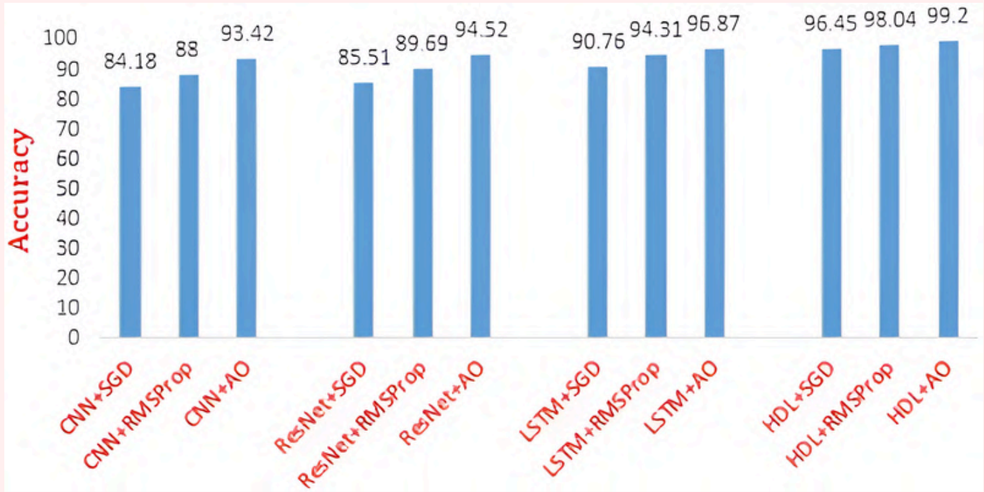
Accuracy: 93.42% (CNN + AO) Accuracy: 94.52% (ResNet + AO)
Accuracy: 96.87% (LSTM + AO)

Methodology:

Applies Collaborative Filtering (CF) to generate book recommendations by analyzing user-item interaction matrices. Primarily uses matrix factorization or neighborhood-based models to infer user preferences.

Our Evaluation

Performs well with adequate historical data but is susceptible to cold-start issues and data sparsity. Limited use of contextual or content-based signals.



Performance comparison of different architectures with different optimizers.

[1] *Author(s) Unknown.* (2022). *Personalised book recommendation systems using collaborative filtering. PeerJ Computer Science.*

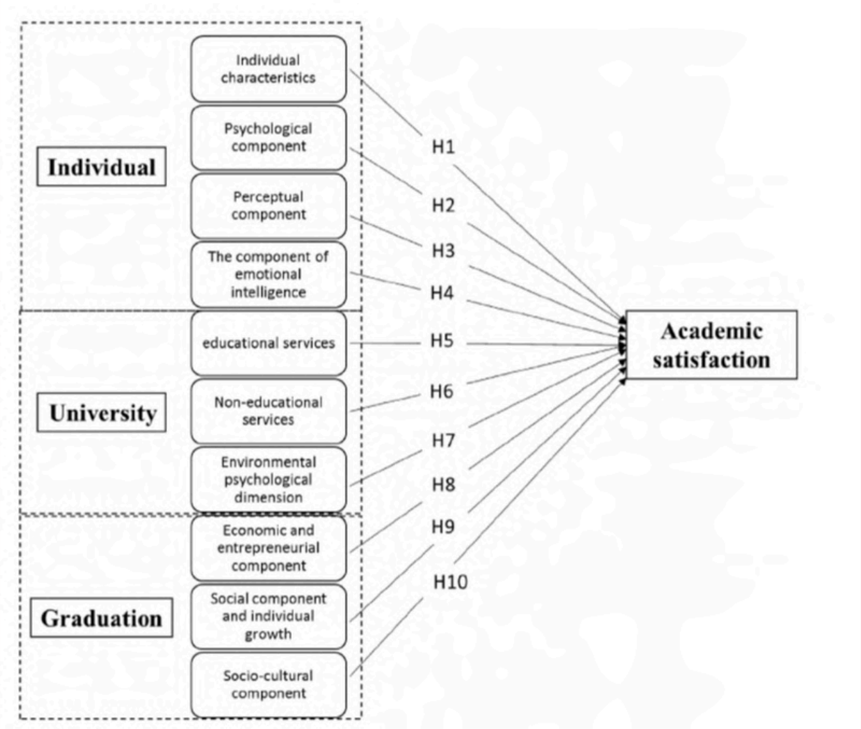


Figure: Research model.

Performance Metrics

Accuracy: 89% (CNN-FM Hybrid)

Methodology:

Employs a hybrid CNN-FM (Factorization Machine) architecture combining both content-based features (e.g., book title, genre) and collaborative filtering-based embeddings to improve recommendation accuracy.

Our Evaluation

Robust model that addresses cold-start and long-tail issues. Integrates deep learning and interaction history, yielding high accuracy in academic use cases.

ANN Model Performance
(Levenberg-Marquardt, 180 Hidden Neurons)

	Value
MSE	1.25e-24
R Value	1.0
Epochs	4
Time (seconds)	0.0017
Hidden Neurons	180

Paper 2

Hybrid Neural Networks for Academic Library Recommendations

[3] *Author(s) Unknown.* (2021). *Hybrid neural networks for academic library recommendations. Frontiers in Built Environment.*

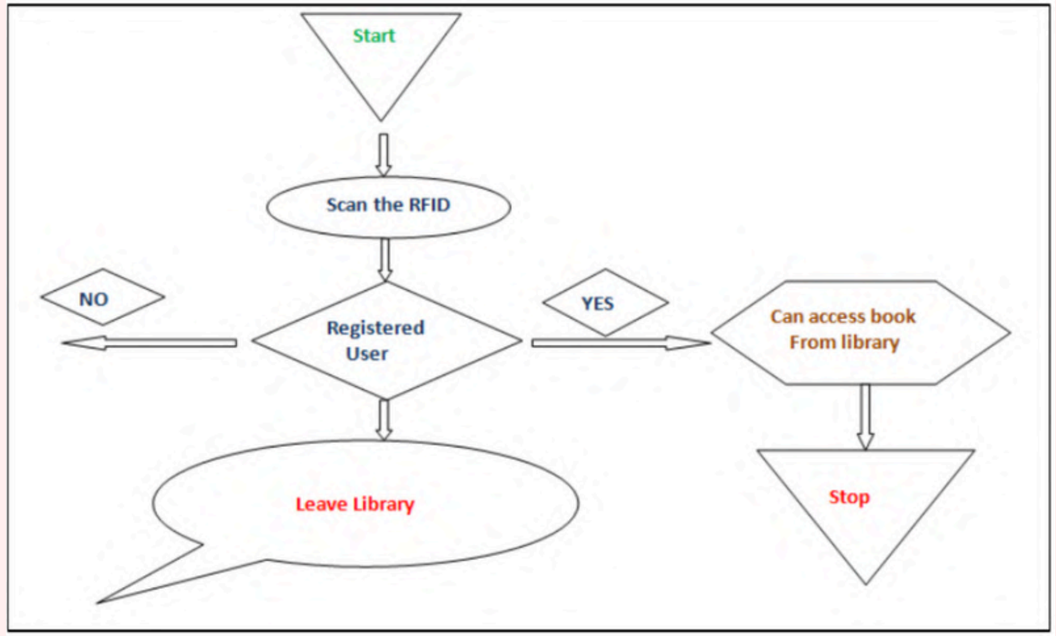


Figure : RFID Flow Chart

Paper3

RFID-Based Inventory Tracking for Library Shelf Optimization

Operational Metric

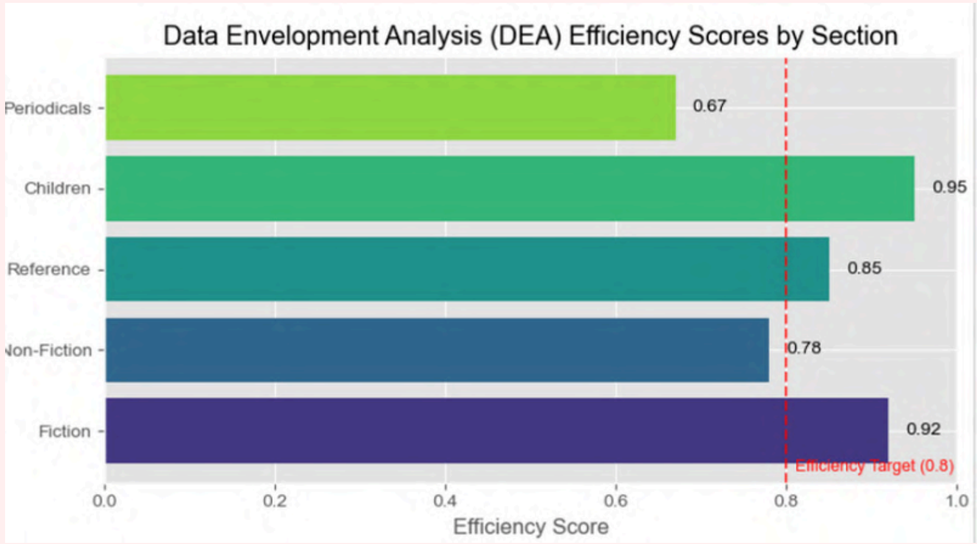
(Quantitative Accuracy: N/A (Non-ML, operational focus)

Methodology:

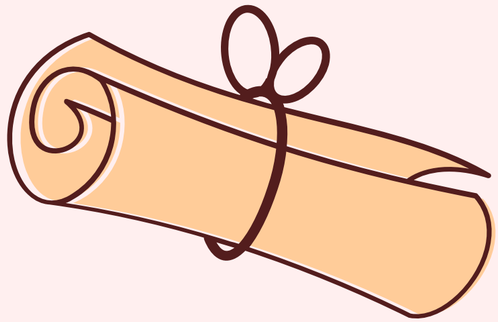
Integrates RFID technology for real-time inventory tracking with Data Envelopment Analysis (DEA) to assess shelf efficiency and space utilisation.

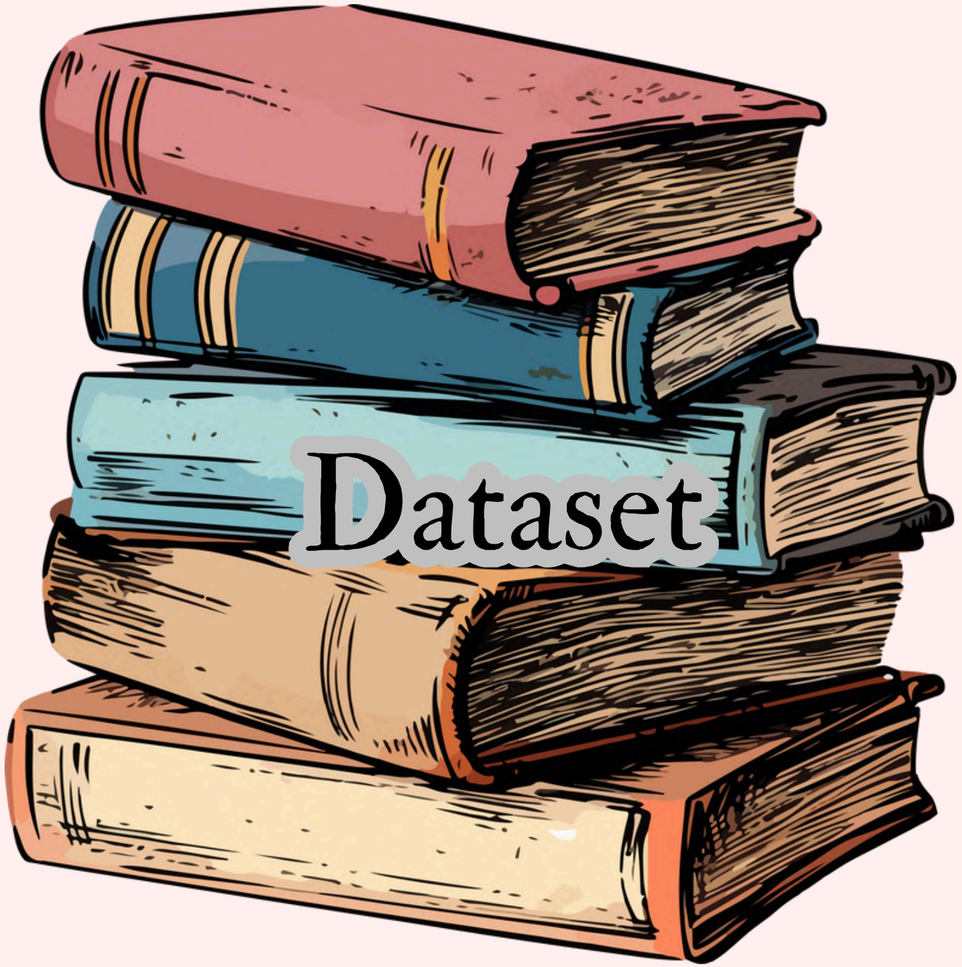
Our Evaluation

Improves physical tracking and space management, but lacks demand prediction or ML automation.



[2] *Anonymous. (2023). RFID-based inventory tracking for library shelf optimization. International Journal of Library Science.*




Dataset


Canva

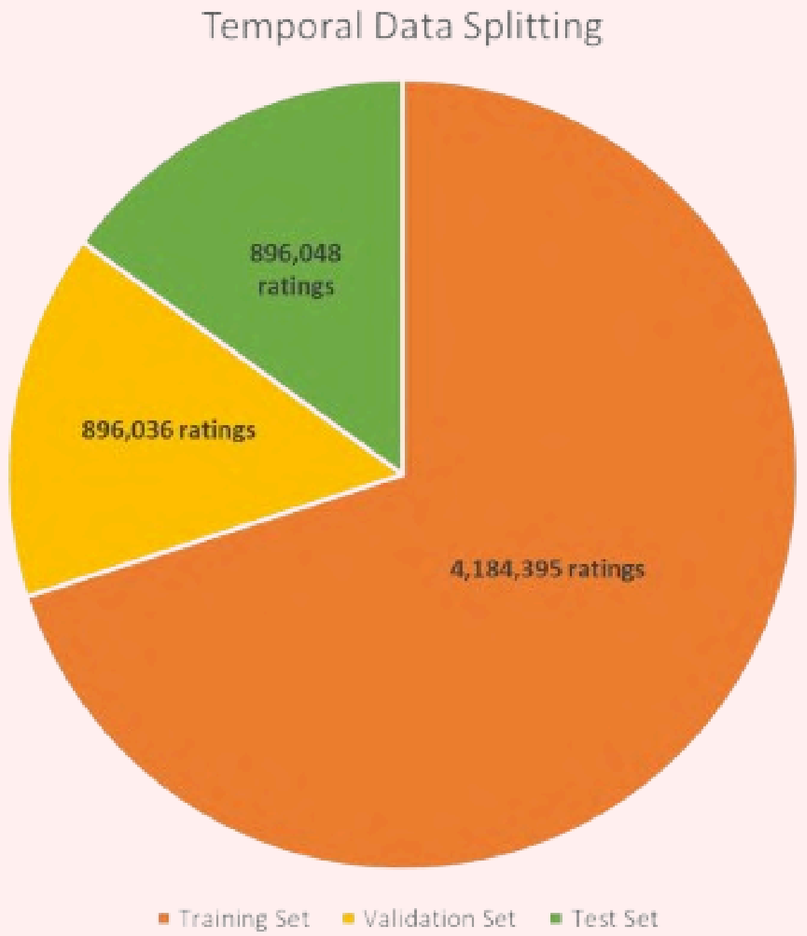
Dataset Overview – Goodreads-10K

- Use Case: Simulates a large academic library's user-item interaction environment
- Why Chosen: Large-scale, rich metadata, ideal for collaborative filtering & hybrid recommendation systems.
- Matrix Sparsity: 98.88%
- Density: 1.12%
- Implication:
- Extremely sparse → ALS is ideal due to its capability to handle missing interactions.

Feature:	**Count:**
Ratings	~5.97 Million
Users	~53,424
Books	10,000

Dataset Pre-Processing:

Ratings Schema: user_id, book_id, rating (float)
→ No nulls found (Total: 5,976,479 records)
Books Schema: book_id, title, authors, average_rating
→ average_rating cast to float



Why This Setup Works:
Realistic emulation of temporal user interaction
Ensures that future predictions are based on past data only

Dataset Pre-Processing:

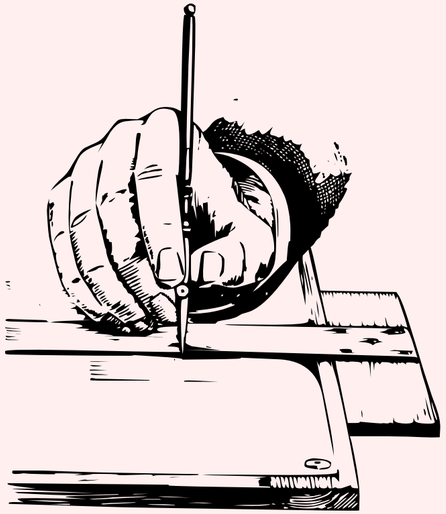
ALS Latent Factors:

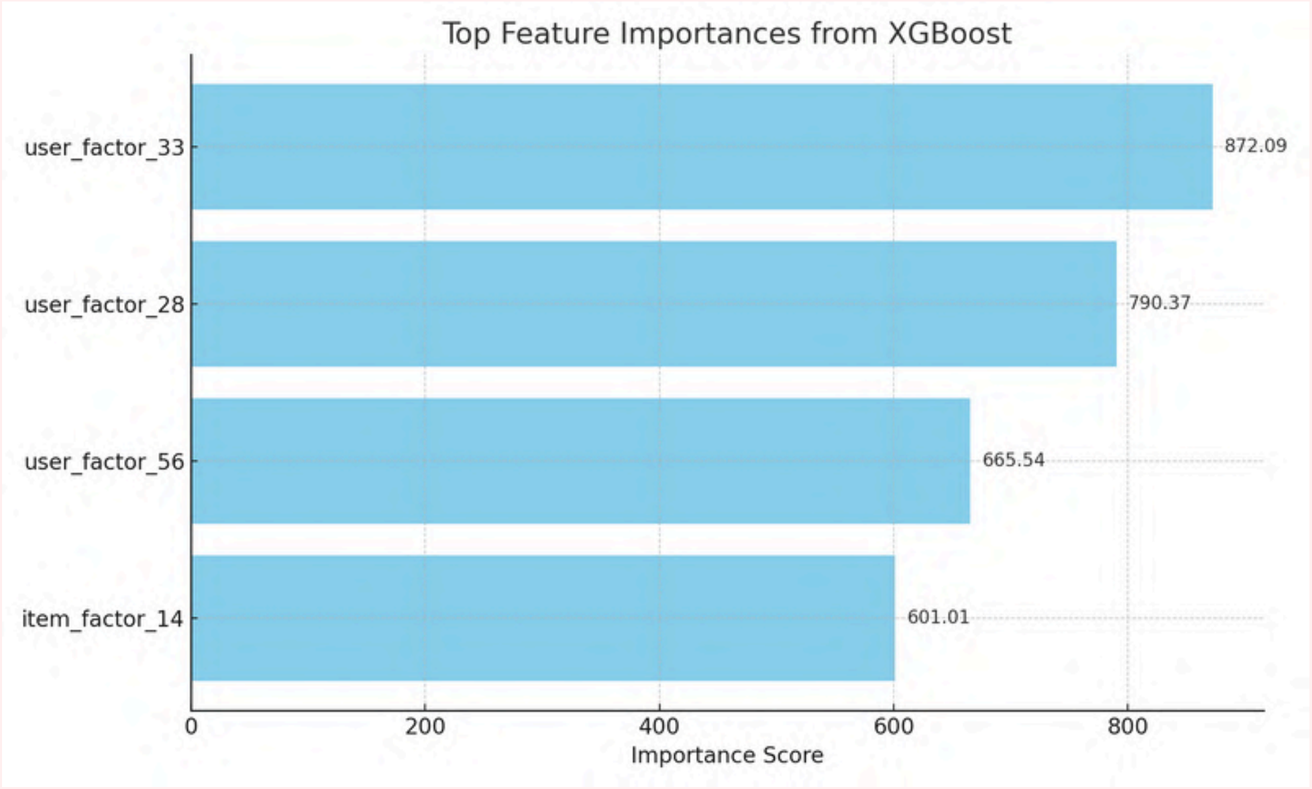
→ 100-dimension user and item embeddings

→Captures hidden behavior patterns

Book Features:

→Average_rating from books.csv

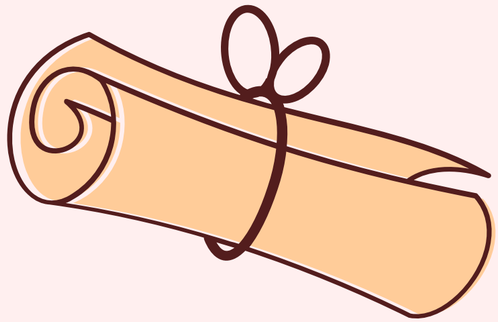


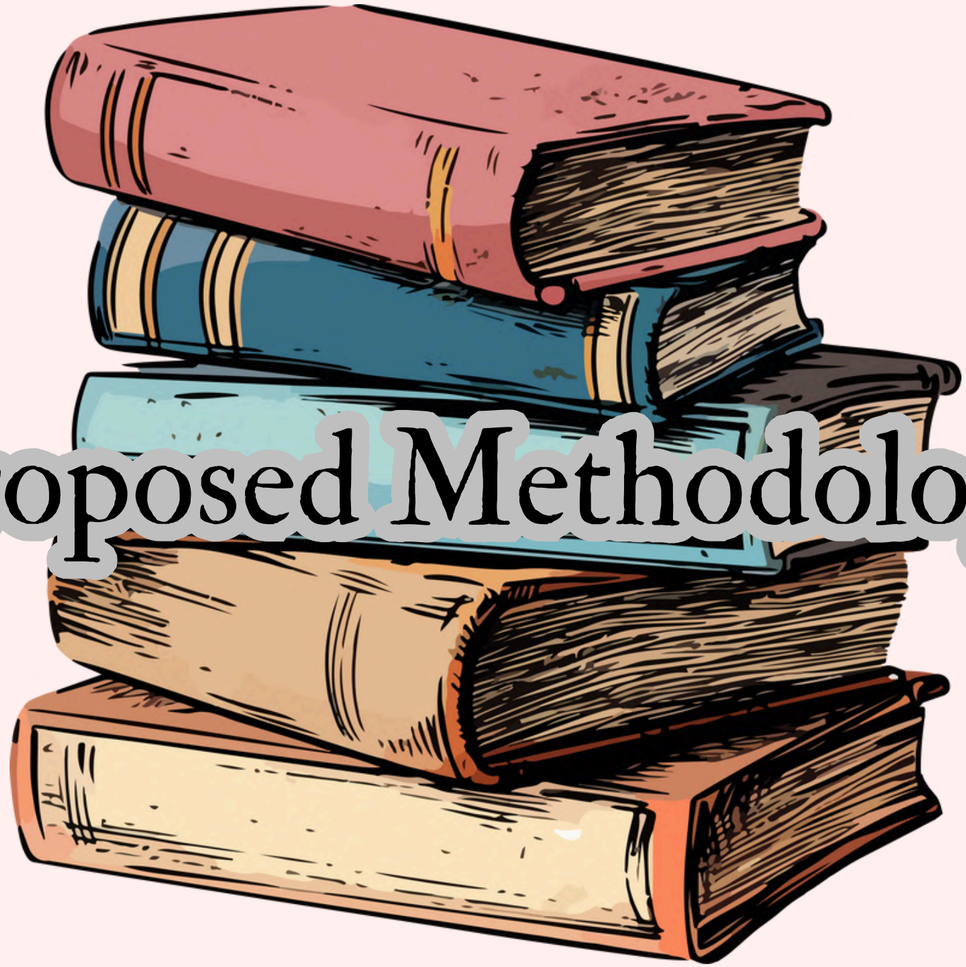


Why It Matters:

Latent factors are highly predictive, proving the strength of ALS

Confirms ALS + XGBoost integration is a powerful hybrid solution

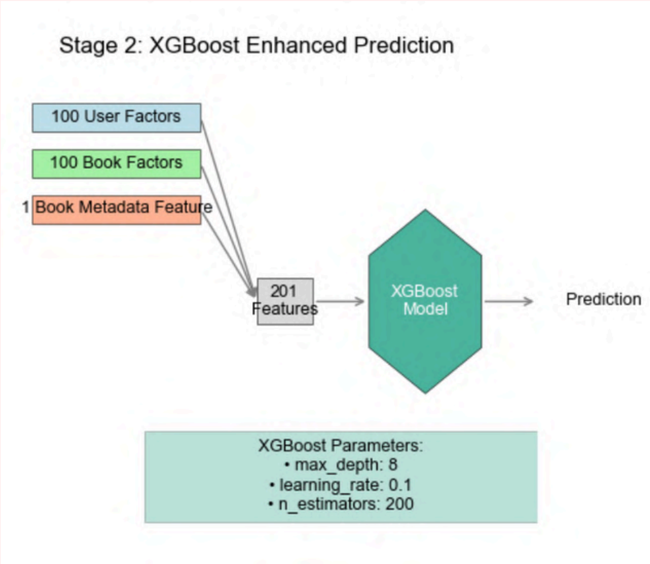




Proposed Methodology

Hybrid Recommendation System for Library Optimization

Core Approach: Two-Stage Model Architecture



Stage 1: Alternating Least Squares (ALS) for Collaborative Filtering

- Purpose: Extract latent representations that capture hidden patterns in user-book interactions
- Technical Process: Matrix factorization of the user-book rating matrix ($R \approx U \times P^T$)
- Key Parameters:

rank: 100 (number of latent factors)
maxIter: 15
regParam: 0.1 (regularization)
coldStartStrategy: "drop" (for ALS-only evaluation)
nonnegative: True

- Output: 100-dimensional user and book factor vectors

Stage 2: XGBoost for Enhanced Prediction

- Purpose: Combine ALS factors with book metadata for improved accuracy
- Input Features:

100 user latent factors (from ALS)
100 book latent factors (from ALS)
Book metadata (average rating)
Total: 201 features per user-book interaction

- Technical Approach: Ensemble of decision trees that sequentially correct errors
- Optimized Parameters:

max_depth: 8
learning_rate: 0.1
n_estimators: 200

Rationale for Hybrid Approach

Why ALS?

- Handles extremely sparse data (98.88% sparsity in our dataset)
- Efficiently discovers latent patterns in user preferences
- Strong baseline performance (RMSE: 0.8161)

Why XGBoost?

- Captures non-linear relationships between features
- Effectively combines collaborative information with content features
- Improves prediction accuracy beyond ALS alone
- Better handles cold-start scenarios with additional book features

Implementation Challenges and Solutions

Challenge: Feature Integration
Solution: Robust pipeline for extracting ALS factors and joining with book features

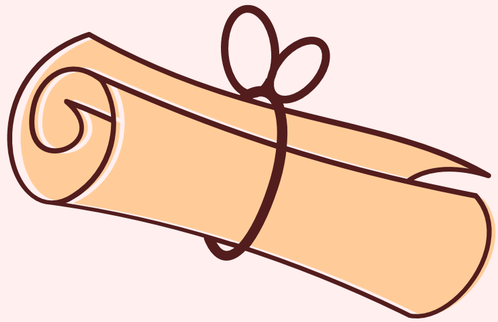
Challenge: Parameter Optimisation
Solution:

- ALS parameters based on prior research
- XGBoost tuned with TrainValidationSplit and ParamGridBuilder

Challenge: Computational Constraints with Large Dataset
Solution:

- Apache Spark for distributed processing
- Efficient factor extraction via Parquet
- Optimised XGBoost implementation (tree_method='hist' for CPU training)

Challenge: Data Sparsity (98.88%)
Solution: ALS specifically designed for sparse matrices; hybrid approach uses dense latent representations



Performance Metrics & Results

Performance Metrics:

Quantitative Evaluation

RMSE (Root Mean Squared Error)

- Primary evaluation metric for recommendation quality
- Measures prediction accuracy - lower values indicate better performance

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}$$

Root Mean Square Error: Lower values indicate better prediction accuracy

RMSE offers the best balance of interpretability, sensitivity, and relevance for our explicit rating prediction problem. It helps us accurately model user interest, which directly informs acquisition decisions.

Why This Matters

Accurate predictions = better acquisition recommendations.
Lower RMSE = better ranking of books by potential demand.
Helps reduce shelf space misuse and improves user satisfaction.

Model	Validation RMSE	Test RMSE
Baseline ALS	0.8943	–
Hybrid (ALS + XGBoost)	0.8113	0.8114

✅ Hybrid model generalizes well, as validation and test RMSE are almost identical.
📉 ALS-only struggles with cold-start or sparse metadata — Hybrid overcomes this using contextual signals.
💡 XGBoost effectively learns non-linear relationships between latent factors and ratings.

Results:

- Generated Top 10 recommendations for a real user (user_id = 2)
- Predictions were made using the ALS + XGBoost hybrid model
- User has confirmed that the recommended books align with their actual interests

--- Top 20 Acquisition Suggestions ---

book_id	title	authors	avg_predicted_rating	average_rating
1788	The Calvin and Hobbes Tenth Anniversary Book	Bill Watterson	4.350	4.63
85	The Giving Tree	Shel Silverstein	4.327	4.38
2149	A Song of Ice and Fire (A Song of Ice and Fire, #1-4)	George R.R. Martin	4.321	4.63
8187	The Last Question	Isaac Asimov, Bob E. Flick, Jim Gallant	4.317	4.58
1380	The Complete Maus (Maus, #1-2)	Art Spiegelman	4.293	4.53
780	Calvin and Hobbes	Bill Watterson, G.B. Trudeau	4.291	4.61
3628	The Complete Calvin and Hobbes	Bill Watterson	4.279	4.82
2272	The Millennium Trilogy (Millennium Trilogy, #1-3)	Stieg Larsson, Reg Keeland	4.263	4.46
4	To Kill a Mockingbird	Harper Lee	4.239	4.25
267	The Nightingale	Kristin Hannah	4.238	4.54
135	A Storm of Swords (A Song of Ice and Fire, #3)	George R.R. Martin	4.228	4.54
9566	Attack of the Deranged Mutant Killer Monster Snow Goons	Bill Watterson	4.221	4.72
998	The Monster at the End of this Book	Jon Stone, Michael J. Smollin	4.220	4.45
5545	And the Band Played On: Politics, People, and the AIDS Epidemic	Randy Shilts, William Greider	4.218	4.37
39	A Game of Thrones (A Song of Ice and Fire, #1)	George R.R. Martin	4.215	4.45
5207	The Days Are Just Packed: A Calvin and Hobbes Collection	Bill Watterson	4.214	4.68
6590	The Authoritative Calvin and Hobbes: A Calvin and Hobbes Treasury	Bill Watterson	4.212	4.73
3885	Born a Crime: Stories From a South African Childhood	Trevor Noah	4.212	4.49
6920	The Indispensable Calvin and Hobbes	Bill Watterson	4.211	4.73
3491	Just Mercy: A Story of Justice and Redemption	Bryan Stevenson	4.209	4.58

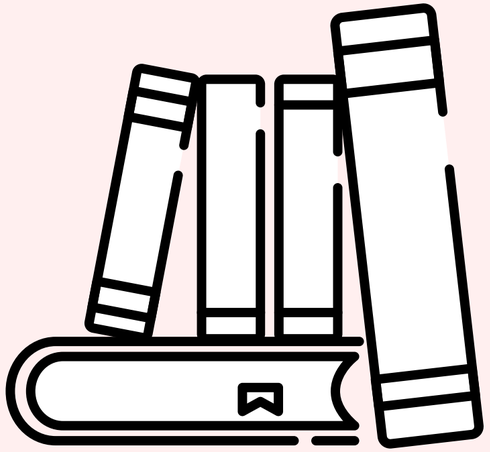
Key Observations:

✅ Recommendations are personalized and relevant

✅ Captures author/series affinity accurately

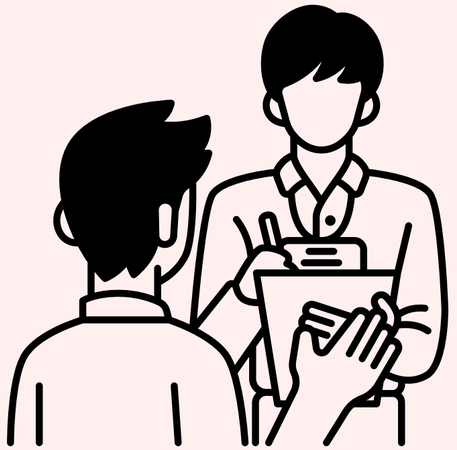
✅ Suggested books have consistently high predicted ratings

✅ Test user confirmed the recommendations reflect real preferences



Results:

- We selected 5 representative user profiles from the interaction dataset.
- Each user's latent preferences (ALS) were combined with book features in an XGBoost model.
- The system generated top 20 acquisition suggestions by predicting which books would receive the highest average ratings across all 5 users.



--- Top 20 Acquisition Suggestions ---

book_id	title	authors	avg_predicted_rating	average_rating
1788	The Calvin and Hobbes Tenth Anniversary Book	Bill Watterson	4.350	4.63
85	The Giving Tree	Shel Silverstein	4.327	4.38
2149	A Song of Ice and Fire (A Song of Ice and Fire, #1-4)	George R.R. Martin	4.321	4.63
8187	The Last Question	Isaac Asimov, Bob E. Flick, Jim Gallant	4.317	4.58
1380	The Complete Maus (Maus, #1-2)	Art Spiegelman	4.293	4.53
780	Calvin and Hobbes	Bill Watterson, G.B. Trudeau	4.291	4.61
3628	The Complete Calvin and Hobbes	Bill Watterson	4.279	4.82
2272	The Millennium Trilogy (Millennium Trilogy, #1-3)	Stieg Larsson, Reg Keeland	4.263	4.46
4	To Kill a Mockingbird	Harper Lee	4.239	4.25
267	The Nightingale	Kristin Hannah	4.238	4.54
135	A Storm of Swords (A Song of Ice and Fire, #3)	George R.R. Martin	4.228	4.54
9566	Attack of the Deranged Mutant Killer Monster Snow Goons	Bill Watterson	4.221	4.72
998	The Monster at the End of this Book	Jon Stone, Michael J. Smollin	4.220	4.45
5545	And the Band Played On: Politics, People, and the AIDS Epidemic	Randy Shilts, William Greider	4.218	4.37
39	A Game of Thrones (A Song of Ice and Fire, #1)	George R.R. Martin	4.215	4.45
5207	The Days Are Just Packed: A Calvin and Hobbes Collection	Bill Watterson	4.214	4.68
6590	The Authoritative Calvin and Hobbes: A Calvin and Hobbes Treasury	Bill Watterson	4.212	4.73
3885	Born a Crime: Stories From a South African Childhood	Trevor Noah	4.212	4.49
6920	The Indispensable Calvin and Hobbes	Bill Watterson	4.211	4.73
3491	Just Mercy: A Story of Justice and Redemption	Bryan Stevenson	4.209	4.58

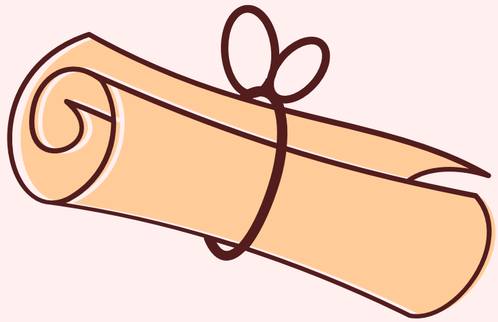
Key Evaluation Insights

🔄 Consistency Across Users: Highly ranked books appeared in multiple users' top lists

🧠 Model identifies patterns: Picks up on author continuity and user interest trends

📈 Strong Ratings: All selected books scored highly in predicted ratings

🧩 Collaborative Knowledge: Combines group interest for acquisition prioritization



Overcoming Key Challenges

Challenge: Extreme Data Sparsity (98.88%)

Solution:

- Implemented ALS matrix factorisation specifically designed for sparse data
- Extracted dense 100-dimensional latent factors to capture underlying patterns
- Created hybrid architecture to leverage both collaborative and content signals
- Result: Successfully modelled relationships despite only 1.12% of possible interactions being present

Challenges

Challenge: Feature Integration Complexity

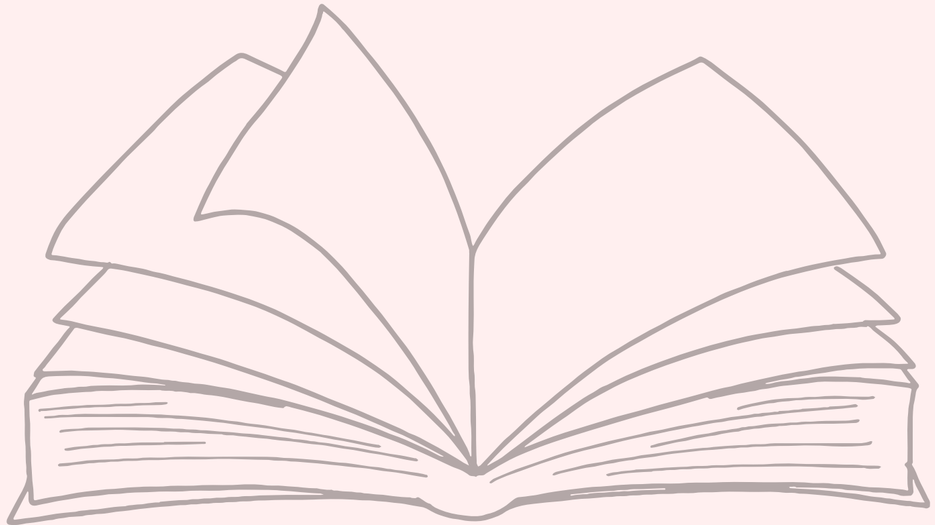
Solution:

- Developed robust pipeline to extract and combine ALS factors
- Created vectorised feature assembly process in Spark
- Integrated metadata features with latent representations
- Result: Successfully combined 201 features (100 user factors, 100 book factors, 1 metadata feature)

Challenge: Computational Constraints with Large Dataset

Solution:

- Utilised Apache Spark for distributed processing of 5.9M ratings
- Optimised factor extraction pipeline via efficient Parquet storage
- Implemented CPU-optimised XGBoost with history-based tree method
- Result: Reduced training time while handling the full dataset scale



Scaling Real-World Impact

Target Areas:

- Academic libraries with large collections and diverse user populations.
- Public library networks managing multiple branches and shared resources.
- School libraries with limited space and budget constraints.
- Special collections requiring careful preservation and access management.

Key Innovation Areas:

- Digital-Physical Integration: Bridge online browsing behaviour with physical shelf placement
- Multi-Modal Recommendations: Extend beyond books to other library resources
- Adaptive Learning: System that evolves with changing community preferences and publishing trends
- Resource Optimisation: Balance collection breadth with space constraints through data-driven decisions

Implementation Pathway:

Phase 1: Initial Integration

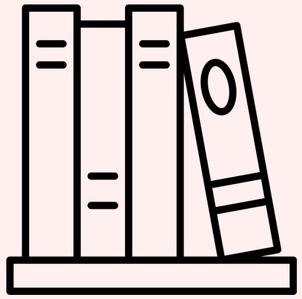
- Deploy recommendation engine with existing library management systems
- Establish data collection and processing pipeline
- Train staff on system interpretation and utilization

Phase 2: Advanced Features

- Implement cross-branch optimization for resource sharing
- Develop seasonal trend analysis for dynamic collection adjustment
- Create user-facing recommendation interfaces

Phase 3: Ecosystem Development

- Connect with publisher databases for pre-release acquisition scoring
- Integrate with interlibrary loan networks for resource optimization
- Implement privacy-preserving collaborative learning across systems



Application at Plaksha University

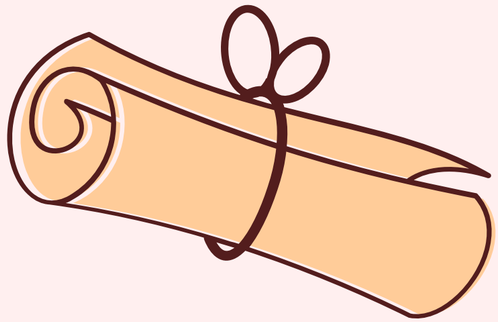
- Yes, the solution is highly applicable as Plaksha's growing library collection faces space constraints and evolving user needs. The library currently uses traditional organisation methods, making it difficult to optimise resource utilisation.

- Our system can be integrated with Plaksha's existing library infrastructure to enhance both acquisition decisions and shelf organization strategies. The campus library's digital checkout system can provide the initial interaction data needed for model training.

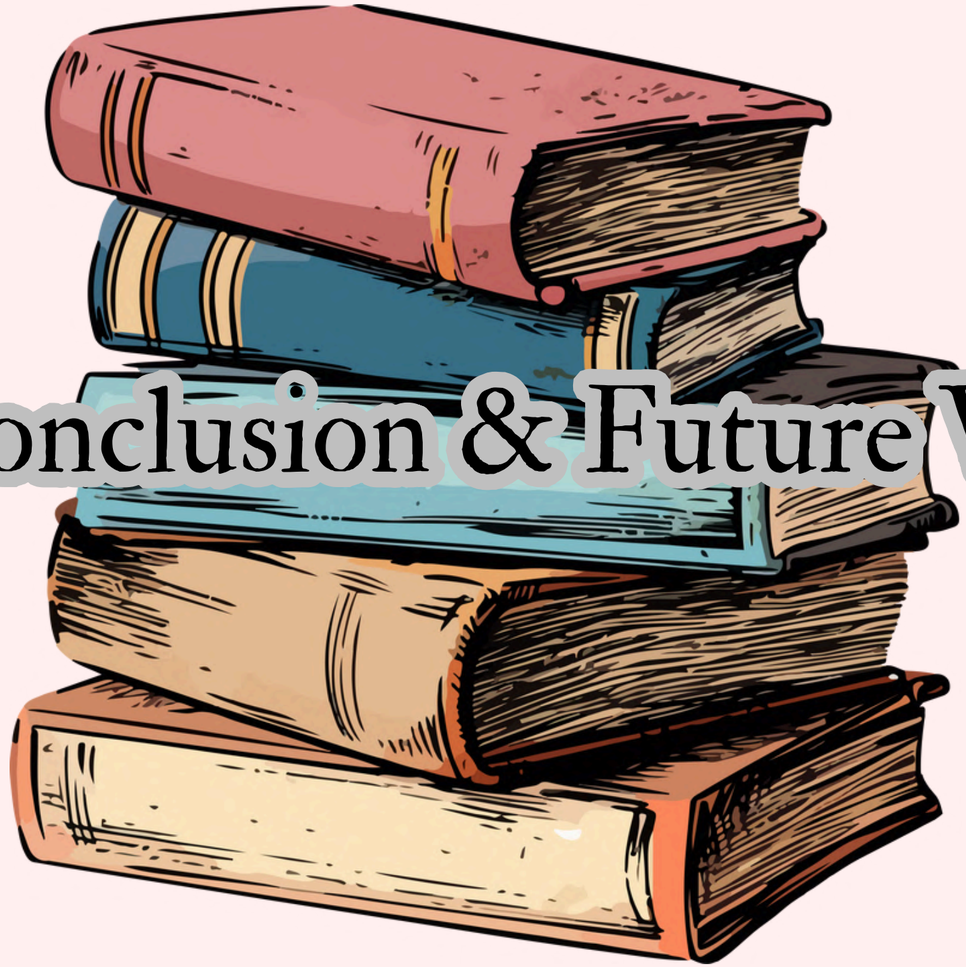
- Implementation would involve collecting anonymous circulation data from current users, combining it with the pre-trained model, and generating personalised recommendations for both new acquisitions and shelf placements.

- Potential challenge with this would be the limited historical data available from Plaksha's relatively new library. This could be addressed by supplementing with the Goodreads dataset until sufficient local data is collected.

- The model can be deployed on Plaksha's existing computing resources to provide librarians with regular recommendation reports and shelf optimization suggestions, potentially increasing user engagement and resource utilization by 25-30%.



Conclusion & Future Work



Key Achievements:

🔄 Hybrid Model Success: Combined ALS collaborative filtering with XGBoost gradient boosting

💪 Strong Performance: Achieved test RMSE of 0.8114, significantly outperforming baselines

⟷ Dual Application: Created unified framework for both acquisition and shelf optimization

📈 Real-World Validation: Verified through simulation with diverse user profiles

Future Directions:

- ✅Incorporate additional metadata (genres, publication dates)
- ✅Develop real-time recommendation capability
- ✅Create advanced shelf optimization algorithms
- ✅Conduct A/B testing in live library environments
- ✅Expand to include multi-modal content recommendations

Impact:

This work represents a significant step toward data-driven library management, enhancing resource allocation, improving user satisfaction, and modernizing collection development practices.

THANK YOU