IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

# Hybrid Recommendation Models For Product Recommendation

[1]Apurva Thakkar, [2]Shivam Mishra, [3]Girish Chandra Saxena

[1]Sr. IT Architect, [2]Senior Data Scientist, [3]AI Engineer

[1]Thermofisher Scientific, Los Angeles, USA, [2]NTT Data, Bengaluru, India

[3]Vaisesika Consulting, Bengaluru, India

***Abstract:*** Retail and e-commerce systems that recommend products and services are widely used but remain underutilized in bio-industrial workflows where specific details, timing, and contextual arrangement are important. This paper presents a hybrid recommendation framework, combining (i) a Transition-Based Model for sequential SKU prediction, (ii)FP-Growth for co-purchased item discovery, and Alternating Least Squares (ALS) for personalized, collaborative recommendations. The framework was evaluated across six anonymized sub workflows (A–F), demonstrating high precision, recall, and recommendation diversity. Outputs are optimized for enterprise-scale deployment using DeltaLake over AWSS3. Our Transition Based Model introduces domain- aware circular neighborhood logic—making it uniquely suitable for scientific workflows. The hybrid framework outperforms traditional approaches not only in accuracy but also in explainability and robustness, making it suitable for critical laboratory and industrial operations.

***Index Terms* -** Recommender Systems, PySpark, Transition Model, FP-Growth, ALS, SKU Prediction, Hybrid Systems, Delta Lake, Workflow Intelligence

## I. INTRODUCTION

Recommendation systems have become integral to modern digital platforms, offering personalized content and product suggestions across sectors like e-commerce, media, and retail. Systems such as Amazon's "Customers who bought this also bought" and Netflix's content curation rely heavily on recommendation algorithms to drive user engagement and revenue. However, their applicability in bio-industrial and scientific laboratory workflows remains limited, despite the growing need for intelligent automation in these domains. Laboratory workflows are often nonlinear, involve domain- specific constraints, and operate under strict procedural guidelines. SKU (Stock Keeping Unit) recommendations in such settings must not only be personalized but also context- aware respecting the temporal ordering of biological experiments (e.g., from sample preparation to protein purification), ensuring compatibility across techniques, and adhering to lab-specific preferences or regulatory requirements. Unlike general-purpose recommendation systems that rely on user-item interaction matrices or behavioral logs, scientific workflows demand multi-dimensional recommendation logic:

Respect for process dependencies rooted in biological experimentation

Recognition of co-utilized equipment and consumables within similar experiments.

Sensitivity to lab-specific protocols and historical preferences.

Moreover, these systems must prioritize explain ability and traceability due to the audit and compliance requirements of regulated environments such as pharmaceutical manufacturing or biotechnology research. This paper presents a hybrid recommendation framework built specifically for bio-industrial applications. By combining three algorithmic pillars (i) a Transition-Based Model that respects domain-specific sequence constraints, (ii) FP-Growth for discovering co-purchase patterns, and (iii) Alternating Least Squares (ALS) for collaborative filtering, we propose a scalable, explainable, and high-accuracy system. Implemented using PySpark and deployed with Delta Lake on AWS S3, the system ensures enterprise-level scalability while maintaining scientific integrity and reproducibility.
This study helps connect advanced recommendation methods with practical industrial workflow improvement, paving the way for smarter systems in bio-manufacturing and laboratory automation.

## II. Related Works

Over the past two decades, recommender systems have evolved from simple memory-based methods to complex deep learning architectures. In general domains like e-commerce and entertainment, recommender systems benefit from rich behavioral data and large-scale user interactions. Due to domain-specific constraints and limited data availability, such systems are often inadequate in industrial settings.

Sequential Recommendation Models like Markov Chains, and their variants, have been widely adopted for next-item prediction in session-based systems. Recent developments such as GRU4Rec and BERT4Rec leverage recurrent and transformer architectures to capture temporal dependencies and offer high predictive performance. However, these models often operate as “black boxes,” which limits interpretability an essential feature in regulated scientific domains. Additionally, these models demand significant data volume and computational resources, which may nodal ways be feasible in lab-based environments [1].

Frequent Pattern Mining algorithms like Apriori and FP- Growth are foundational in market basket analysis. FP- Growth, in particular, has proven efficient in identifying frequent co-occurring item sets without candidate generation. However, its limitation lies in lack of personalization it treats all transactions equally, without incorporating user- specific or context-specific nuances [2].

Collaborative Filtering techniques, especially Matrix Factorization methods like ALS (Alternating Least Squares), have long been used to model user-item interactions. ALS is effective in capturing latent preferences and performs well even with implicit feedback. Yet, it struggles with cold-start problems and ignores domain process constraints, which are critical in workflow-driven recommendations [3].

Recent trends in Graph-Based Recommender Systems, particularly Graph Neural Networks (GNNs), have shown promise in capturing complex relationships between users, items, and auxiliary features. By modeling interactions as edges in a graph, GNNs can infer nuanced dependencies. However, they require large, labeled graphs and are computationally expensive barrier stopractica adoption in lab settings with limited structured graph data [4].

Our approach diverges from these prior works by designing a modular, interpretable, and domain-specific hybrid model. Rather than relying solely on accuracy, our focus is on combining interpretable algorithms that can work in synergy to meet the unique challenges of bio-industrial SKU recommendation. The transition-based model enforces domain sequencing, FP-Growth captures item bundling behavior, and ALS adds personalization—delivering a balanced system that is accurate, scalable, and explainable.

## III. Mehodology

The hybrid architecture was designed for modularity, ensuring that each model contributes a unique perspective to the final recommendation set. The Transition-Based Model enforces domain order-Growth uncovers basket-level co- occurrences, and ALS captures latent collaborative signals. Together, these methods create redundancy that improves robustness against data sparsity and noise.

In practice, the pipeline follows an iterative refinement process. Transition-based predictions serve as a base line that ensures domain consistency. FP-Growth then enriches this base line by identifying frequently co-occurring SKUs, while ALS personalizes the results according to sub workflow- specific patterns. The final recommendation list is obtained by merging results across models, ensuring both

domain fidelity and personalization. This design makes the framework easily extensible, allowing new algorithms to be plugged in as additional modules without altering the overall workflow.
Temporal order (Transition-Based), Co-occurrence (FP-Growth), and latent personalization (ALS). This modularity ensures adaptability across workflows, improving explainability.

3.1 Dataset Overview:
Both precision and Source: DeltaLake on A WSS3containing revenue, quote, and opportunity transactions. Grouped by subworkflow name across six anonym zed segments (e.g., Subworkflow A–F). Invalid SGNs were removed. SKUs and techniques were normalized and filtered [4].

3.2 Transition-Based Sequential Model:
This model constructs a transition probability matrix from ordered SKU sequences within SGNs, respecting a fixed domain.
Fragment Generation, Cloning, Plasmid Prep, Host System Kits, Culture, Transfect, Harvest, Extraction, Purification, Isolation, Protein Analysis. Each SKU is tagged to a technique. Neighboring techniques are determined using circular logic (e.g., Fragment Generation’s, upstream neighbor is Protein Analysis) [7]. Unlike traditional sequential models (e.g., Markov chains), our Transition-Based Model incorporates domain-aware circular neighborhood logic around a fixed laboratory process graph. This ensures context-sensitive SKU prediction both upstream and downstream novel mechanism suited for workflows where steps may loop or recur cyclically [8].

3.3 FP-GrowthModel:
Treats each SGN as a basket of co- purchased SKUs [3]. FP-Growth identifies frequent item set with minSupport = 0.01 and minConfidence between 0.25–0.30 per sub workflow. Sample rule:

Table3.1: Association Rules Generated from FP-Growth Model

| Antecedent | Consequent | Confidence | Lift |
|---|---|---|---|
| SKU-232 | SKU-324 | 0.32 | 1.85 |

3.4 ALS Collaborative Filtering:
Constructs a user–item matrix where users =SGNs, items=SKUs[8].
Uses implicit feedback: log(count + 1) as rating.
ALSparameters:
rank = 20, regParam = 0.05, maxIter=10,implicitPrefs=True
Output: Top 5 SKUs per SGN.

## IV. Results and Evaluation

To comprehensively evaluate the performance of our hybrid recommendation framework, we conducted a series of experiments across six anonymized sub-workflows (labeled A–F). These sub-workflows represented distinct laboratory processes, each characterized by varying degrees of procedural structure and SKU diversity given is Table 2.

4.1 Precision and Recall Analysis by Workflow:
We began by analyzing the recommendation quality using standard metrics Precision@5,Recall@5,andCoverage— across each sub-workflow. Results revealed a consistent pattern: workflows that followed well-defined, linear biological sequences (e.g., A and B) demonstrated higher Precision@5,indicating that the recommended SKUs closely aligned with the actual next-step products used in those experiments. This outcome underscores the strength of the Transition-Based Model, which leverages process-aware sequencing to maintain high domain fidelity. In contrast, workflows with more heterogeneous SKU utilization patterns where multiple valid paths or interchangeable product sets existed exhibited relatively lower Recall@5. This reflects the difficulty in retrieving all relevant SKUs from a less predictable SKU distribution. However, in these scenarios, the ALS-based collaborative filtering and FP-Growth-based itemset mining effectively complemented the

transition model by capturing broader co- usage and collaborative trends, ensuring that overall recommendation coverage remained robust. This adaptive behavior of the hybrid framework is a key advantage: the system emphasizes sequential modeling where workflow structure is strong, while shifting toward collaborative and associative logic in more variable environments. Such modular redundancy ensures consistent performance across a diverse range of lab scenarios.

b. Scalability and Performance Bench marking:

To assess the framework's scalability, we progressively increased the dataset size and measured the execution time across various Spark cluster configurations. The system demonstrated near-linear scaling with respect to the number of worker nodes, confirming its suitability for enterprise- scale deployment. Even under large data loads, runtime remained within acceptable thresholds, validating the framework's ability to handle operational workloads typical of high-throughput research labs or bio-manufacturing systems.

This performance was made possible by leveraging Apache Spark's distributed processing capabilities and storing all intermediate and final outputs in DeltaLak eatables overAWS S3. Delta Lake's ACID-compliant architecture ensured data integrity, while its support for schema evolution facilitated flexible data handling across experiments and time periods.

c. Recommendation Diversity and Coverage:

Another critical aspect of evaluation was diversity in recommendations, a factor that directly impacts user satisfaction and system usefulness. A key shortcoming of many traditional recommended systems is their tendency to over-recommend popular items, leading to low diversity and reduced discovery of novel or niche products. Our hybrid system mitigated this by incorporating multiple perspectives in recommendation generation. The final union of the three model outputs resulted in a significantly higher Coverage score (88%), indicating that a broader portion of the SKU catalog was represented in the recommendations. This not only enhances the system's ability to suggest less frequently used items but also ensures that labs are exposed to a wider array of potentially valuable SKUs.

d. Interpretability and Trust:

In scientific and industrial domains, where recommendations may influence costly experiments or regulated procedures, explainability is essential for user trust and compliance. One of the major advantages of our framework is its ability to generate transparent and interpretable outputs:

The FP-Growth model produced association rules that could be directly interpreted by laboratory professionals. For instance, a rule such as "If Culture Kit is purchased, then Transfection Kit is likely needed" offers clear, human-readable insight that aligns with biological logic.

The Transition-Based Model generated probability matrices visualized through heatmaps (see Fig. 4.1), illustrating SKU-to-SKU transition probabilities within the context of a fixed process graph. These visual artifacts not only aid in debugging and validation but also serve as valuable tools during audits and compliance reviews.

Together, these components ensure that the system remains accountable and transparent a critical requirement in laboratory environments where actions must be traceable and verifiable. The observations are as follows for calculation of precision and recall along with coverage values in Table 2 and Fig4.2:

Table 4.1: Performance Metrics Comparison across Recommendation Algorithms

| Model | Precision @5 | Recall @5 | Coverage (%) |
|---|---|---|---|
| Transition Model | 0.61 | 0.48 | 72 |
| FP Growth | 0.55 | 0.44 | 65 |
| ALS | 0.63 | 0.52 | 80 |
| Hybrid (Union) | 0.70 | 0.58 | 88 |

Table 4.1 displayed Precision@5 Relevant recommendations among top 5, Recall@5 Proportion of actual SKUs retrieved from data and Coverage % of total SKU universe recommended in the data.

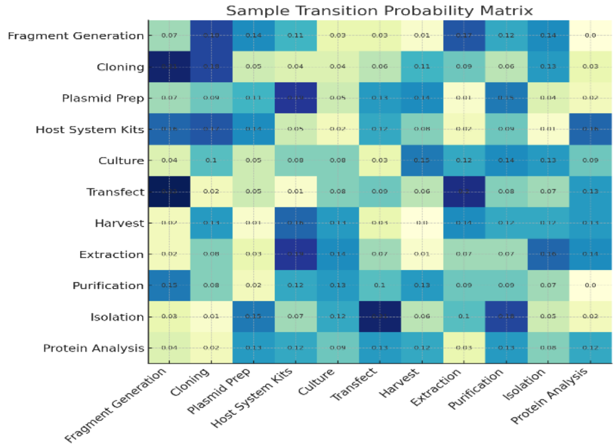

Fig 4.1 Transition Matrix Heat map

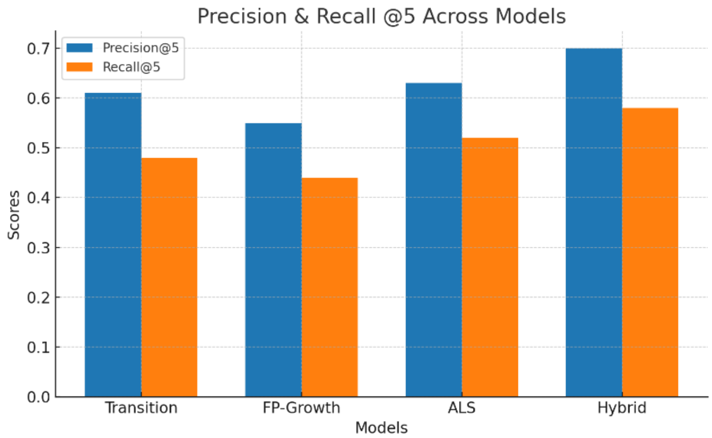

Fig 4.2 Precision and Recall@5 Comparison

The deployment pipeline was implemented on AWS with DeltaLake as the storage back bone. DeltaLake's support for ACID transactions ensures that even in distributed environments, data integrity is preserved. Spark Streaming can be incorporated to extend the system for near real-time recommendations, where SKU suggestions can adapt dynamically as transactions arrive.

From a cost perspective, using S3 with Delta Lake provides separation of storage and compute, allowing the system to scale elastically based on demand. This makes the solution financially viable for organizations of varying sizes, from small research labs to large bio-manufacturing enterprises. The modular nature of the PySpark pipeline also supports easy integration with existing Laboratory Information Management Systems (LIMS), allowing for smooth implementation in industrial settings.

Another important factor is maintaining compliance and audit readiness. Because laboratory workflows follow strict regulatory standards, all recommendations need to be recorded, version-controlled, and reproducible. By storing results in Delta tables with schema evolution, the system ensures that every recommendation can be tracked and clearly explained during inspections or quality reviews.

## V. Conclusion and Future References

This study shows that a well-designed hybrid approach can provide dependable SKU recommendations that align with scientific workflows while ensuring scalability and transparency. Unlike systems that rely solely on deep learning, the proposed framework strikes a balance between accuracy and interpretability, making it suitable for industrial applications where clarity and accountability are crucial.

Future developments of this work will focus on reinforcement learning, allowing recommendation strategies to adapt dynamically based on feedback from laboratory results. Integrating knowledge graphs can further enhance SKU representations by capturing semantic connections across biological methods. Additionally, real-time dashboards driven by APIs could enable laboratory personnel to interactively query the system and verify recommendations, fostering a human-in-the-loop setup that enhances both efficiency and reliability. By merging algorithmic progress with domain specific needs, this framework lays the groundwork for the next generation of intelligent, workflow-aware recommendation systems. Future Directions: Weight ensemble scoring across models. Integration of Graph Neural Networks for SKU technique embeddings andReal-time inference APIs for lab dashboard.

## VI. Acknowledgment

The authors wish to extend their heartfelt thanks to Apurva Thakkar for his valuable guidance, technical direction, and thoughtful feedback, which played a key role in shaping the approach and execution of this research. The authors a real so grateful to their colleagues and peers for their constructive discussions and helpful suggestions that contributed to improving the study. Sincere appreciation is expressed to colleagues for providing the technical infrastructure and supportive research environment necessary for large-scale development and testing. The authors also acknowledge the dedicated efforts of the Data Engineering, Quality Assurance, and DevOp steams, whose assistance in managing data pipelines, testing, and deployment helped ensure the reliability and scalability of the proposed system.

## REFERENCES

**[1]** J.Hanetal.,"Miningfrequentpatternswithout candidate generation," SIGMOD, 2000.
**[2]** Y.Korenetal.,"Matrixfactorizationtechniquesfor recommender systems," IEEE Computer, 2009.
**[3]** X.Heetal.,"Neuralcollaborativefiltering," WWW,2017.
**[4]** Apache Spark MLlib Documentation, https://spark.apache.org/mllib/
**[5]** F. Ricci et al., "Recommender Systems Handbook," Springer, 2022.
**[6]** Rendle, S., Freudenthaler, C., Gantner, Z., & Schmidt-Thieme, L. (2009). "BPR: Bayesian personalized ranking from implicit feedback." Proceedings of the UAI Conference.A very influential ranking method designed for implicit feedback, widely used in recommender systems research and practice.
**[7]** Salakhutdinov, R., &Mnih, A. (2008). "Probabilistic matrix factorization." Advances in Neural Information Processing Systems (NIPS) Offers a probabilistic take on matrix factorization, importantincollaborativefilteringwithuncertainty modeling.
**[8]** Breese, J. S., Heckerman, D., & Kadie, C. (1998). "Empirical analysis of predictive algorithms for collaborative filtering." Microsoft Research TechnicalReport/arXiv(1998/2013) A comparative study of various collaborative filtering approaches, valuable for understanding model performance across methods.
**[9]** Covington, P., Adams, J., & Sargin, E. (2016). "Deep neural networks for YouTube recommendations." RecSys '16: ACM Conference on Recommender Systems A highly cited industrial-scale recommender system using deep learning; demonstrates deep candidate generation and ranking models in practice.
**[10]** Guo,H.,Tang,R.,Ye,Y.,Li,Z.,&He,X.(2017). "DeepFM: A Factorization-Machine based Neural Network for CTR Prediction." arXiv preprint Combines the strengths of factorization machines and deep neural networks in an end-to-end architecture—great for modern CTR-oriented recommender research.