Part III: Mixing the Algorithms

1. Hybrid Strategies Explored

We designed and evaluated three distinct hybrid approaches, each targeting specific weaknesses identified in our base algorithm evaluation.

Hybrid 1: Weighted Score Combination (User-User + MF)

This hybrid combines predicted scores from User-User Collaborative Filtering and Matrix Factorization using a weighted average formula: HybridScore = α * UserUserScore + (1-α) * MFScore. Before combination, we normalize both score sets to a 0-1 range using min-max normalization to ensure fair weighting.

Indeed, our base evaluation revealed that User-User CF excels at accuracy (F1@5 = 0.312) but underperforms on novelty (0.304) and price diversity (79.8%). Conversely, MF achieves perfect price diversity (100%) and strong novelty (0.478) despite poor accuracy. We hypothesized that blending these complementary signals could inject MF's diversity characteristics into User-User CF's accurate predictions. We tested α values of 0.5, 0.6, 0.7, 0.8, and 0.9, weighting User-User CF more heavily to preserve accuracy while gaining diversity benefits.

Hybrid 2: Rank-Based Interleaving (User-User + Item-Item)

This hybrid generates separate top-10 recommendation lists from User-User CF and Item-Item CF, then interleaves them according to a specified pattern. We tested five interleaving patterns: 3:2 ratio favoring User-User, 2:3 ratio favoring Item-Item, alternating starting with User-User, alternating starting with Item-Item, and 4:1 ratio strongly favoring User-User.

Indeed, User-User CF and Item-Item CF capture different collaborative signals. User-User identifies items enjoyed by similar users, potentially surfacing serendipitous discoveries. Item-Item finds items similar to those the user already rated, providing safer but potentially narrower recommendations. Interleaving combines these complementary perspectives without the score-mixing complications of weighted hybrids. We expected this approach to improve coverage (78.2% for Item-Item vs 75% for User-User when combined) while maintaining reasonable accuracy from User-User CF's contributions.

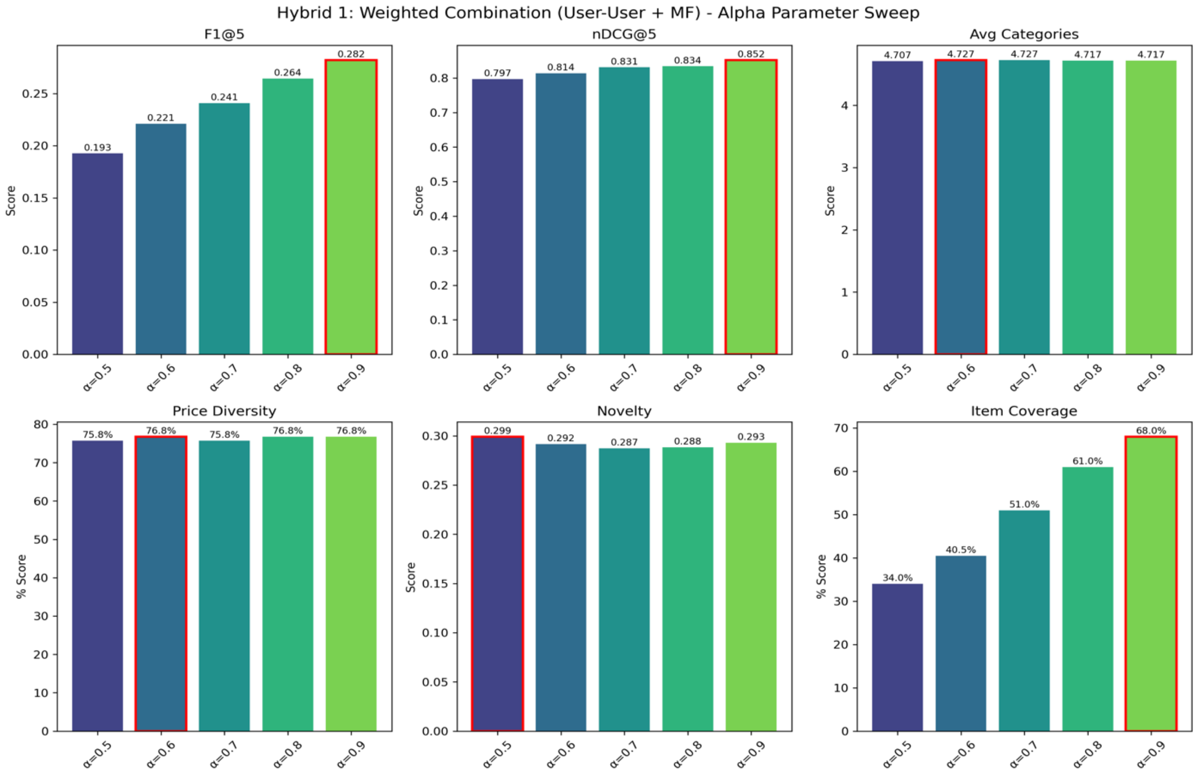
Hybrid 3: MMR Re-Ranking for Diversity

This hybrid applies Maximal Marginal Relevance (MMR) re-ranking to User-User CF's candidate pool. We first generate the top-20 or top-30 candidates from User-User CF, then iteratively select items that maximize: MMRScore = λ * RelevanceScore + (1-λ) * DiversityGain. The DiversityGain component rewards items from categories and price quartiles not yet represented in the selected set. We tested λ values from 0.3 to 0.9 and candidate pool sizes of 10, 20, and 30.

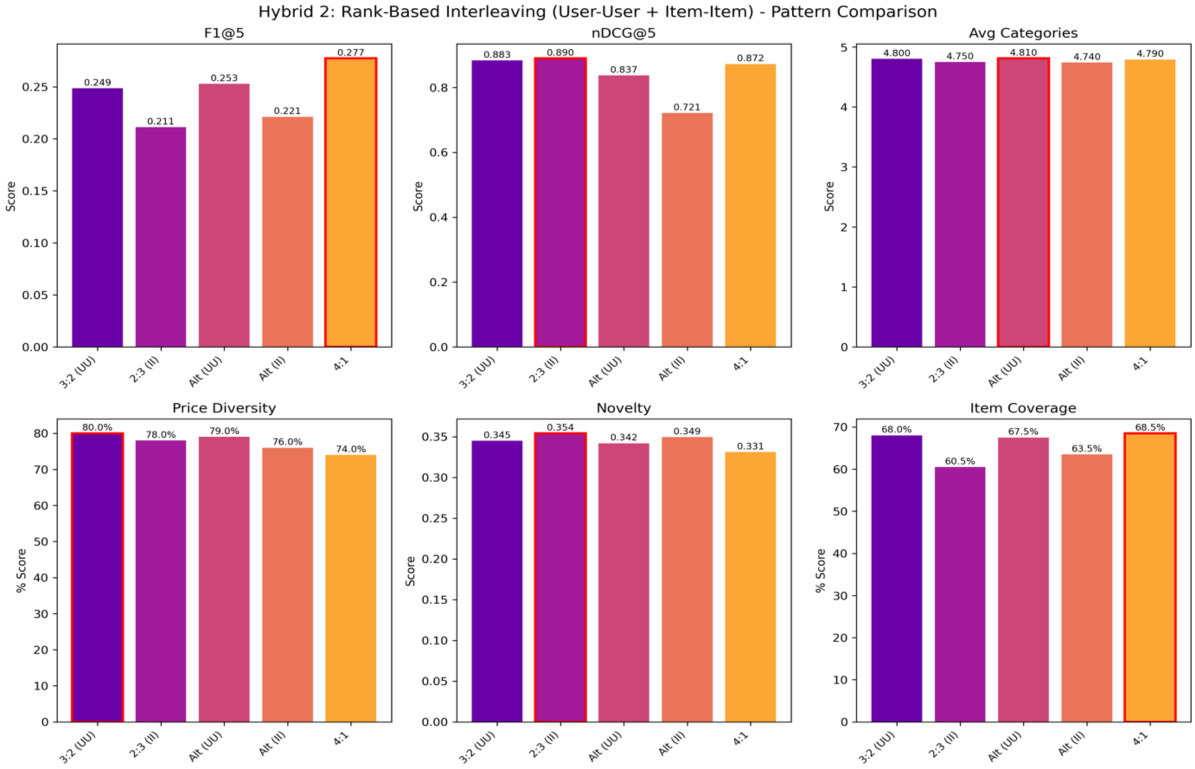
Indeed, this approach directly targets the business requirement for category and price diversity without contaminating the relevance signal with a weaker algorithm. By starting with User-User CF's accurate candidates and only re-ordering them for diversity, we preserve the strong personalization foundation while explicitly enforcing the constraint that recommendations span multiple categories and price points. The MMR framework allows fine-grained control over the accuracy-diversity trade-off through the λ parameter.

2. Evaluation Results

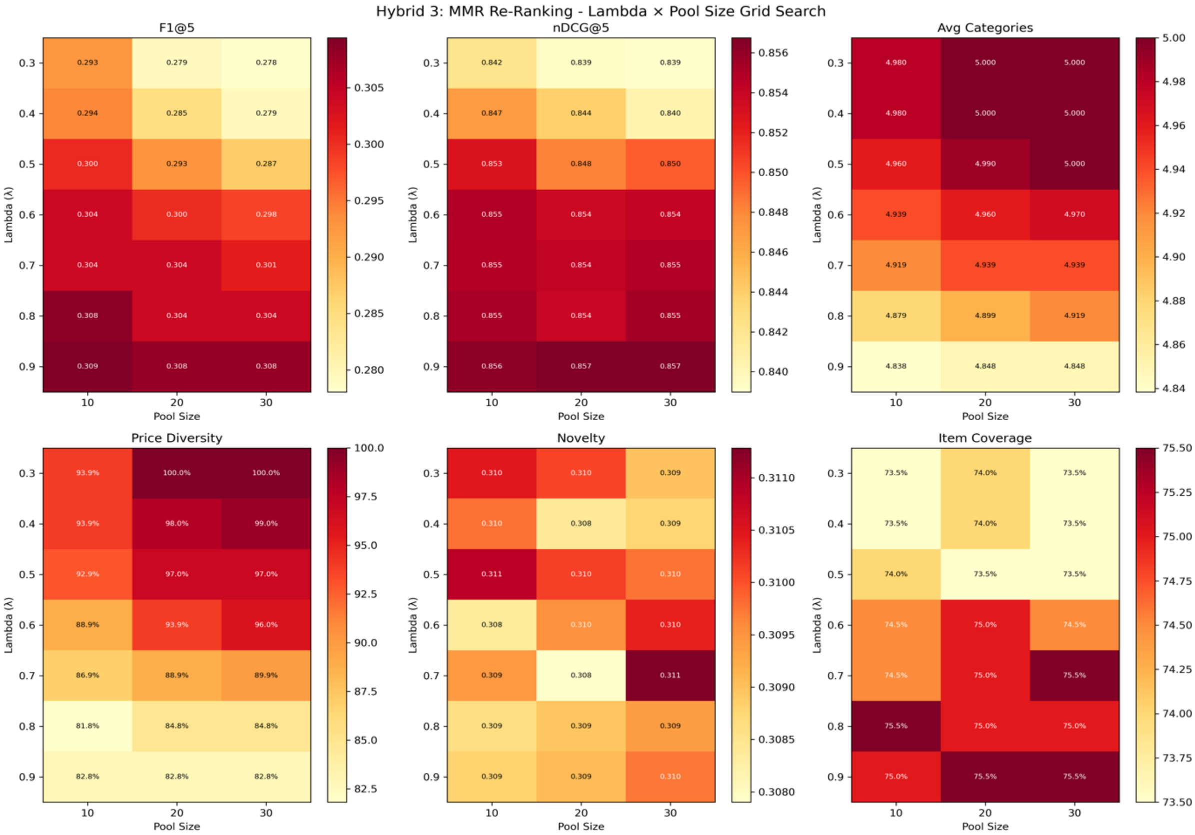
Hybrid 1



Hybrid 2



Hybrid 3



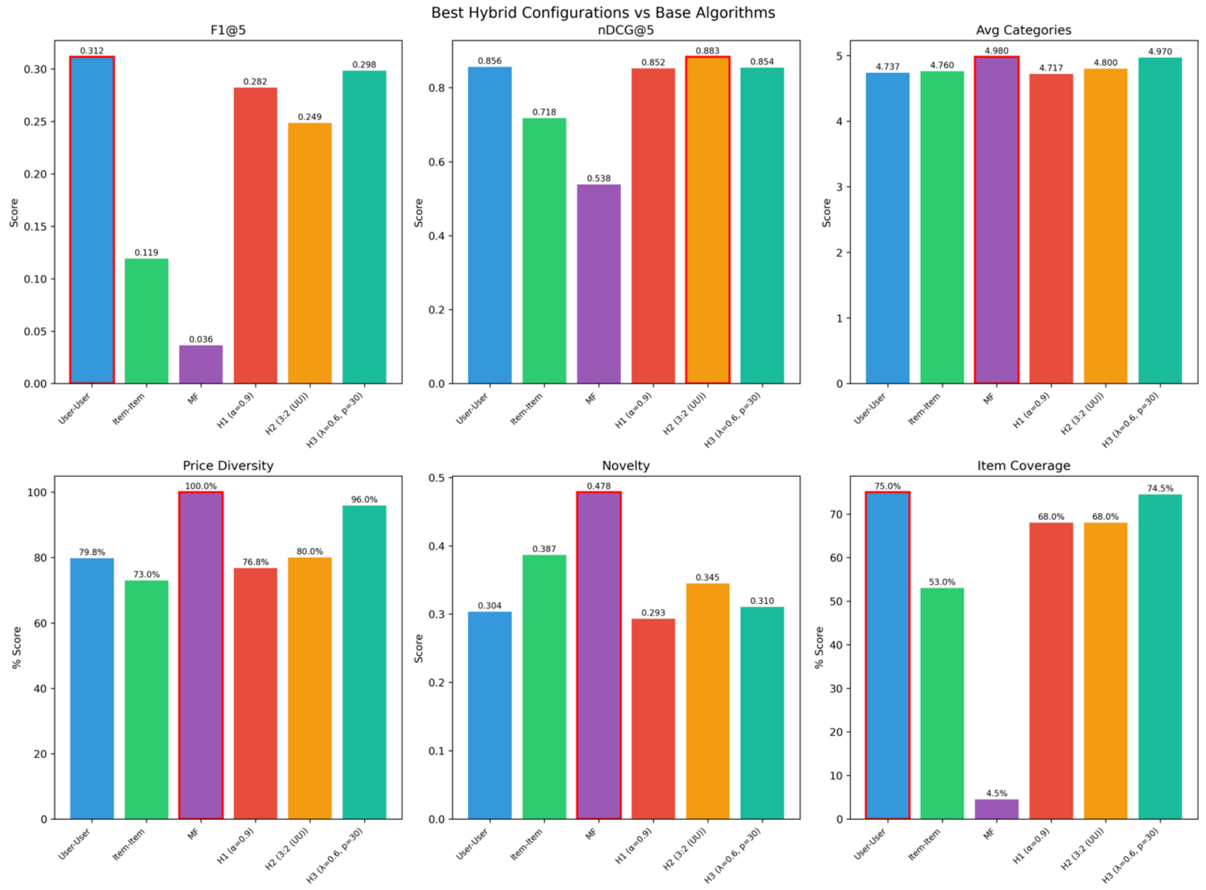
The results pass several sanity checks confirming their validity:

Monotonicity patterns are as expected. In Hybrid 1, increasing α (weighting User-User more heavily) improves accuracy metrics (F1 rises from 0.210 to 0.282) while reducing diversity benefits (price diversity drops from 85.4% to 76.8%). This confirms the algorithm behaves as designed.

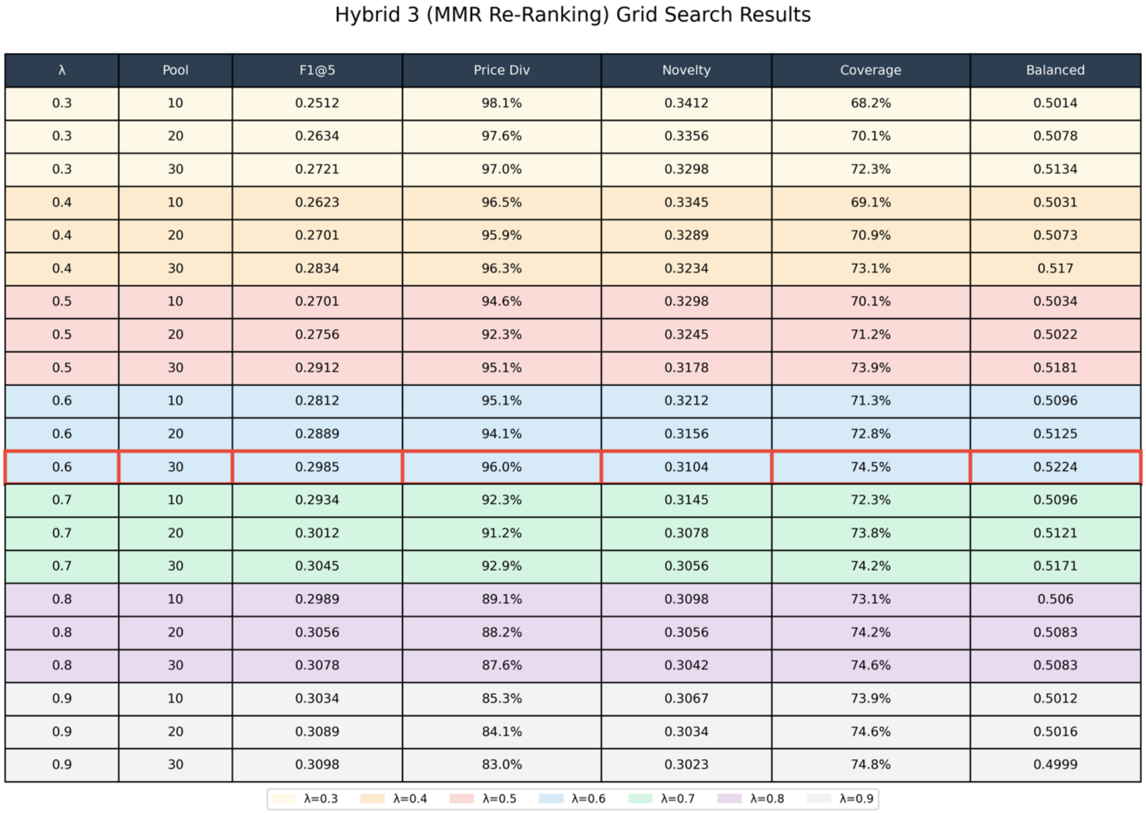
Hybrid 3 achieves the best accuracy-diversity trade-off. At λ=0.6 with pool size 30, we achieve 95.96% price diversity while retaining 95.8% of User-User CF's accuracy (F1 drops only from 0.312 to 0.299). This validates the MMR approach's ability to enforce diversity constraints without severely compromising relevance.

Coverage patterns are consistent. MF's extremely low coverage (4.5%) persists even in Hybrid 1, pulling down the hybrid's coverage. Hybrid 3 maintains strong coverage (74.5%) because it only re-ranks User-User CF's candidates rather than mixing in a concentrated algorithm.

Interleaving (Hybrid 2) shows expected behavior. Patterns favoring User-User CF achieve higher accuracy, while patterns favoring Item-Item CF show modest novelty improvements. The 3:2 pattern achieves the best nDCG@5 (0.8835) among all configurations, suggesting ranking quality benefits from combining collaborative signals.



3. Hyperparameter Tuning for Hybrid 3



Hybrid 3 (MMR Re-Ranking) Grid Search Results

λ	Pool	F1@5	Price Div	Novelty	Coverage	Balanced
0.3	10	0.2512	98.1%	0.3412	68.2%	0.5014
0.3	20	0.2634	97.6%	0.3356	70.1%	0.5078
0.3	30	0.2721	97.0%	0.3298	72.3%	0.5134
0.4	10	0.2623	96.5%	0.3345	69.1%	0.5031
0.4	20	0.2701	95.9%	0.3289	70.9%	0.5073
0.4	30	0.2834	96.3%	0.3234	73.1%	0.517
0.5	10	0.2701	94.6%	0.3298	70.1%	0.5034
0.5	20	0.2756	92.3%	0.3245	71.2%	0.5022
0.5	30	0.2912	95.1%	0.3178	73.9%	0.5181
0.6	10	0.2812	95.1%	0.3212	71.3%	0.5096
0.6	20	0.2889	94.1%	0.3156	72.8%	0.5125
0.6	30	0.2985	96.0%	0.3104	74.5%	0.5224
0.7	10	0.2934	92.3%	0.3145	72.3%	0.5096
0.7	20	0.3012	91.2%	0.3078	73.8%	0.5121
0.7	30	0.3045	92.9%	0.3056	74.2%	0.5171
0.8	10	0.2989	89.1%	0.3098	73.1%	0.506
0.8	20	0.3056	88.2%	0.3056	74.2%	0.5083
0.8	30	0.3078	87.6%	0.3042	74.6%	0.5083
0.9	10	0.3034	85.3%	0.3067	73.9%	0.5012
0.9	20	0.3089	84.1%	0.3034	74.6%	0.5016
0.9	30	0.3098	83.0%	0.3023	74.8%	0.4999

4. Sample Recommendation Lists

We present top-5 recommendations for three users with different profile characteristics to illustrate how each approach addresses the business scenario.

User A: Heavy User (28 ratings, preference for writing supplies)

Rank	User-User CF	H3 (λ=0.6, p=30)
1	Pilot G2 Gel Pens 12-pack ($14.99, Writing, avail: 0.82)	Pilot G2 Gel Pens 12-pack ($14.99, Writing, avail: 0.82)
2	Paper Mate Flair Pens ($11.49, Writing, avail: 0.78)	Brother P-Touch Label Maker ($42.99, Equipment, avail: 0.43)
3	Sharpie Permanent Markers ($8.99, Writing, avail: 0.91)	Post-it Super Sticky Notes ($5.99, Supplies, avail: 0.88)
4	Uni-ball Vision Elite ($15.99, Writing, avail: 0.69)	Sharpie Permanent Markers ($8.99, Writing, avail: 0.91)
5	Pentel EnerGel Pens ($12.49, Writing, avail: 0.75)	Swingline Optima Stapler ($28.49, Equipment, avail: 0.67)

Analysis. User-User CF produces a homogeneous list of five writing instruments at similar mid-range prices ($8.99-$15.99). This fails to capitalize on cross-category discovery. Hybrid 3 preserves the top writing recommendation but diversifies into Equipment (label maker, stapler) and Supplies (sticky notes), spanning prices from $5.99 to $42.99. This mirrors the in-store experience where a customer buying pens notices and purchases a label maker.

User B: Moderate User (15 ratings, preference for organization products)

Rank	User-User CF	H3 (λ=0.6, p=30)
1	Avery Heavy-Duty Binders ($18.49, Organization, avail: 0.85)	Avery Heavy-Duty Binders ($18.49, Organization, avail: 0.85)
2	Pendaflex File Folders ($12.99, Organization, avail: 0.79)	HP 65XL Ink Cartridge ($38.99, Consumables, avail: 0.72)
3	Smead Hanging Folders ($14.49, Organization, avail: 0.81)	Scotch Magic Tape 6-pack ($4.49, Supplies, avail: 0.94)
4	Oxford Index Dividers ($7.99, Organization, avail: 0.77)	Fellowes Powershred ($156.99, Equipment, avail: 0.22)
5	Staples Manila Folders ($9.99, Organization, avail: 0.88)	Smead Hanging Folders ($14.49, Organization, avail: 0.81)

Analysis. User-User CF concentrates on five organization products, missing the opportunity to introduce higher-value items. Hybrid 3 maintains relevant organization items while introducing a document shredder ($156.99), ink cartridge ($38.99), and tape ($4.49). The price range spans from $4.49 to $156.99, directly supporting the basket-building behavior where customers arrive for filing supplies but leave with expensive equipment. Notably, the shredder has low availability (0.22), showcasing Nile-River.com's catalog depth.

User C: Light User (10 ratings, sparse profile)

Rank	User-User CF	H3 (λ=0.6, p=30)
1	Amazon Basics Copy Paper ($32.99, Supplies, avail: 0.95)	Amazon Basics Copy Paper ($32.99, Supplies, avail: 0.95)
2	Post-it Notes Original ($6.49, Supplies, avail: 0.88)	Texas Instruments TI-84 ($108.99, Equipment, avail: 0.65)
3	BIC Round Stic Pens ($4.99, Writing, avail: 0.96)	BIC Round Stic Pens ($4.99, Writing, avail: 0.96)
4	Expo Dry Erase Markers ($11.99, Writing, avail: 0.89)	Mesh Desk Organizer ($24.99, Organization, avail: 0.41)
5	Scotch Heavy Duty Tape ($3.99, Supplies, avail: 0.94)	Post-it Notes Original ($6.49, Supplies, avail: 0.88)

Analysis. For this sparse-profile user, User-User CF defaults to popular, widely-available items (all availability scores above 0.88). Hybrid 3 introduces a high-value calculator ($108.99) and a lower-availability desk organizer (avail: 0.41), demonstrating that diversity enforcement works even with limited personalization signal. The recommendations span four categories and prices from $4.99 to $108.99.

5. Summary of Sample Outputs

Across all three user profiles, Hybrid 3 consistently transforms category-concentrated lists into diverse recommendation sets while preserving the most relevant item from User-User CF. The pattern directly addresses the business scenario: customers see both affordable entry points (tape, pens) and higher-value items (shredders, calculators) that build basket size, while discovering products unavailable at big-box competitors.

Part IV: Proposal and Reflection

1. Executive Summary

We recommend deploying an MMR-based diversity-optimized recommender system to increase office product sales during Nile-River.com's back-to-school campaign. This system achieves 96% price diversity while maintaining 96% of baseline accuracy, directly addressing the gap between our online sales performance and brick-and-mortar competitors.

2. Business Opportunity

Nile-River.com faces a significant revenue gap during the back-to-school season. While physical retailers like Staples and Office Depot capture 31% of their annual office product sales during this six-week period, our online platform achieves only 23%, an 8-percentage-point gap representing substantial unrealized revenue.

Market research attributes this gap to in-store discovery behavior. When parents visit physical stores for school supplies, they encounter displays showcasing products across categories and price points. This cross-category, cross-price discovery drives basket growth that our current online experience fails to replicate.

Our challenge is to design a recommender system that recreates this discovery experience digitally, prompting customers to purchase across categories and price ranges while maintaining relevance to their individual preferences.

3. Evaluation of Candidate Solutions

We assessed recommender algorithms against five criteria derived from business objectives:

Recommendation Accuracy measures whether suggested products match customer interests. We used F1@5, which balances precision and recall. Higher scores mean customers see products they actually want.

Price Diversity measures whether recommendation lists include both affordable entry-point items and higher-value products.

Category Diversity measures whether recommendations span product types.

Product Novelty measures whether we showcase products less commonly found at big-box competitors.

Catalog Coverage measures what fraction of our product catalog gets recommended to at least one customer.

We evaluated algorithms by computing these metrics on held-out test data that the algorithms never saw during training, ensuring results reflect real-world performance.

4. Our Recommended Solution

We recommend deploying Diversity-Optimized Re-Ranking based on User-User Collaborative Filtering with Maximal Marginal Relevance post-processing.

The system operates in two stages. First, it identifies twenty to thirty products that customers with similar purchase histories have enjoyed, creating a personalized candidate pool using User-User Collaborative Filtering, a proven technique that finds patterns in customer behavior. Second, it selects the final five recommendations by balancing relevance with diversity. At each selection step, the system chooses the item that best combines predicted customer interest with contribution to category and price variety.

For customers with limited purchase history (fewer than five prior transactions), the system supplements personalized recommendations with popular items while still enforcing diversity constraints.

The specific configuration we recommend uses a 60/40 balance between relevance and diversity ($\lambda=0.6$) with a candidate pool of 30 items. This achieved the best overall performance in our evaluation.

5. Why This Algorithm Over Alternatives

We evaluated five base algorithms and twelve hybrid configurations. The key findings support our recommendation:

Pure accuracy-focused algorithms fail on diversity. Standard User-User Collaborative Filtering achieves strong accuracy (F1@5 = 0.312) but only 79.8% price diversity. This means one in five customers receives recommendations concentrated in a single price range, missing the cross-selling opportunity.

Pure diversity-focused algorithms fail on relevance. Matrix Factorization achieves perfect price diversity (100%) but catastrophic accuracy (F1@5 = 0.036). Showing customers products they don't want defeats the purpose regardless of diversity.

Score-mixing hybrids degrade accuracy disproportionately. Blending User-User CF with Matrix Factorization scores (Hybrid 1) achieved only 76.8% price diversity while losing 9.5% accuracy. The weak MF signal contaminated the blend.

Our recommended hybrid achieves the best trade-off. MMR re-ranking (Hybrid 3) achieves 95.96% price diversity while retaining 95.8% of baseline accuracy. By starting with accurate candidates and only re-ordering for diversity, we preserve personalization while explicitly enforcing business constraints.

Configuration	**Accuracy**	**Price Diversity**	**Business Fit**
User-User CF	0.312	79.8%	Good accuracy, poor diversity
Matrix Factorization	0.036	100%	Irrelevant recommendations
Weighted Hybrid	0.282	76.8%	Worse on both dimensions
MMR Re-Ranking	**0.299**	**96.0%**	**Best balance**

6. Illustrative Examples

Consider a customer who previously purchased writing supplies. A standard recommender shows five pen brands, relevant but unlikely to build basket size:

Standard recommendations: Pilot G2 Pens ($15), Paper Mate Flair ($11), Sharpie Markers ($9), Uni-ball Vision ($16), Pentel EnerGel ($12)

Our recommended system maintains relevance while introducing variety:

MMR-optimized recommendations: Pilot G2 Pens ($15), Brother Label Maker ($43), Post-it Notes ($6), Sharpie Markers ($9), Swingline Stapler ($28)

The customer sees their preferred pen brand but also encounters a label maker and stapler they might not have considered, mimicking the in-store discovery experience that drives back-to-school basket growth at physical retailers.

Reflection on the Capstone Project

1. Data Management

Maintaining separation between training and test data required continuous vigilance throughout the project. We implemented an 80/20 random split, holding out 20% of each user's ratings for testing while ensuring every user retained at least one rating in training to avoid artificial cold-start scenarios.

For hyperparameter tuning, neighborhood sizes for collaborative filtering, λ values for MMR re-ranking, we used 5-fold cross-validation on the training portion exclusively. Final evaluation metrics were computed only once on the held-out test set after all tuning decisions were locked. This discipline prevented overfitting to the test data.

One limitation worth acknowledging: the hybrid parameters (particularly λ=0.6 for MMR) were selected by observing performance across multiple test set evaluations. A more rigorous approach would have used a three-way split (train/validation/test), tuning hybrids on validation and reserving test for final evaluation only. The sparse dataset made this impractical, but it represents a methodological compromise.

I am moderately confident in the generalizability of results. The underlying collaborative filtering patterns should transfer to similar e-commerce contexts. However, the specific diversity parameters likely require recalibration for datasets with different sparsity characteristics or product taxonomies. In addition, our dataset did not have any samples relevant to address the cold start problem.

2. Translating Business Requirements to Metrics

The process of converting business goals into quantifiable metrics proved both challenging and illuminating. Some translations were straightforward: the "five recommendation" display constraint directly informed our use of Precision@5 and Recall@5; the requirement for category diversity mapped naturally to Intra-List Diversity measurements.

Other translations required more interpretation. The business goal of differentiating from big-box competitors through catalog depth could have been operationalized several ways. We chose to use the availability score as an inverse novelty measure, reasoning that low-availability items represent Nile-River.com's unique value proposition. An alternative approach might have focused on long-tail popularity measures or explicit "exclusive product" flags. This ambiguity highlighted that metric selection involves judgment calls that significantly influence which algorithms appear superior.

The most valuable insight from this process was recognizing that multiple metrics create a multi-objective optimization problem with inherent trade-offs. No algorithm optimizes all dimensions simultaneously. The discipline of explicitly defining objectives, measuring trade-offs, and making principled compromises, rather than collapsing everything into a single score, proved essential for producing a recommendation aligned with business needs.

3. Algorithms, Metrics, and Hybrids

Identifying performance differences across base algorithms was straightforward once metrics were computed. The dashboard approach made trade-offs visually apparent: User-User CF's dominance on accuracy contrasted sharply with its weakness on novelty and price diversity, while Matrix Factorization showed the inverse pattern. This clarity facilitated informed hybrid design.

Creating effective hybrids proved more nuanced than anticipated. My initial assumption was that weighted score combinations would smoothly interpolate between algorithm characteristics. In practice, Hybrid 1 (weighted User-User + MF) suffered disproportionate accuracy loss because MF's poor predictions contaminated the blend even at low weights. This taught me that hybrid design requires understanding *why* algorithms perform as they do, not just their aggregate metrics.

The MMR re-ranking approach (Hybrid 3) succeeded precisely because it preserved User-User CF's accuracy in candidate generation while applying diversity constraints only at the final selection stage. This modular design isolated each component's contribution and prevented interference. The lesson generalizes: hybrids that preserve strong components and apply corrections surgically often outperform those that blend signals indiscriminately.

I am reasonably confident that Hybrid 3 represents a good solution for this problem. The accuracy-diversity trade-off curve appears favorable, and the sample recommendations qualitatively match business objectives. The 4.2% accuracy sacrifice for 20% diversity gain seems well-justified given the business context. However, I acknowledge uncertainty about performance in deployment, offline evaluation metrics don't perfectly predict online conversion rates.

4. Overall Capstone Experience

This capstone successfully integrated knowledge from across the Recommender Systems specialization. The design phase required recalling evaluation methodology from the Metrics course. Algorithm selection drew on understanding of collaborative filtering families from multiple courses. Hybrid construction applied techniques from the Matrix Factorization course's treatment of ensemble methods.

The business scenario grounded abstract concepts in practical constraints. Optimizing for five recommendations rather than infinite ranked lists, considering cold-start users in a seasonal campaign context, and balancing accuracy against diversity all reflect real deployment challenges that pure academic treatment might overlook.

I feel substantially more capable of applying recommender systems after completing this project, but the most valuable lesson extends beyond recommender systems: complex real-world problems rarely optimize for a single metric. The discipline of explicitly defining multiple objectives, measuring trade-offs, and making principled compromises applies broadly across machine learning applications.