HYBRID CLUSTERING AND RECOMMENDATION SYSTEM FOR E-COMMERCE CUSTOMER PERSONALIZATION
FINAL PROJECT REPORT

Chien-Wei Weng & Ke Chen
Nicolas Perion-Quémeneur & Zihan Yang
CentraleSupélec, Université Paris-Saclay
MSc Data Sciences and Business Analytics
chien-wei.weng@student-cs.fr
ke.chen@student-cs.fr
nicolas.perion-quemeneur@student-cs.fr
zihan.yang@student-cs.fr

ABSTRACT

E-commerce grocery platforms face the challenge of personalizing recommendations across diverse customer behaviors. We investigate whether customer segmentation improves recommendation quality compared to global models. Using the Instacart dataset (206,209 users, 49,688 products) (Instacart, 2017), we engineered behavioral features addressing right-skewed distributions in purchase frequency and basket size, including RFM metrics, temporal patterns, and category preferences. K-means clustering (K=5, selected via silhouette and Davies-Bouldin indices over K $\in$ [3,7]) identified five distinct segments: Power Users, Bulk Shoppers, Routine Snackers, Alcohol Enthusiasts, and Household Essentials. We compared three recommendation approaches—Baseline (popularity), Collaborative Filtering (SVD), and Hybrid (CF + Content-Based)—in both global and segment-specific configurations. Rigorous temporal splitting (training on all but last two orders, validation on N-1, testing on N) prevented data leakage. Statistical testing (paired t-tests, n=2,000 stratified users) showed segment-specific Collaborative Filtering and Hybrid models significantly outperformed global counterparts across all metrics ($p < 0.05$), with Hybrid F1@20 improving by 59.5% ($p < 0.001$). Segment-specific Baseline showed no significant improvement ($p > 0.05$), indicating segmentation benefits personalized models but not popularity-based approaches. Results demonstrate that behavioral segmentation measurably enhances recommendation quality in sparse, high-dimensional grocery transaction data.

1 INTRODUCTION

Online grocery retail presents unique recommendation challenges due to high product variety, sparse user-item interactions, and heterogeneous purchase patterns ranging from routine replenishment to exploratory shopping (Miguel et al., 2023). Traditional global recommendation models assume uniform user behavior, potentially overlooking systematic differences in shopping habits, budget constraints, and category preferences. We hypothesize that explicit customer segmentation—grouping users by behavioral similarity—can improve recommendation accuracy by tailoring predictions to distinct shopping profiles.

This study addresses two research questions: (1) Can unsupervised clustering identify interpretable customer segments from transactional features? (2) Do segment-specific recommendation models outperform global models in predicting future purchases? We focus on the Instacart dataset, a large-scale grocery transaction log with 206,209 users and 49,688 products. The user-item interaction matrix exhibits extreme sparsity: with 206,209 $\times$ 49,688 $\approx$ 10.2 billion possible user-product pairs but only 11.6 million unique interactions in training data, the matrix density is approximately 0.11%. Additionally, raw transactional features display right-skewed distributions requiring careful preprocessing.

Our approach integrates feature engineering, clustering, model training, and evaluation through four phases. Phase 0 establishes temporal data splitting and feature engineering: each user's last order serves as the test set, the second-to-last as validation, and all prior orders for training—ensuring no future information leaks into model building. We construct behavioral features (RFM metrics, purchase statistics, temporal regularity, department/aisle preferences) from training data only. Phase 1 applies K-means clustering, evaluating K $\in$ [3, 7] using silhouette and Davies-Bouldin indices, ultimately selecting K=5 for interpretability and cluster quality. We exclude DBSCAN due to its sensitivity to density thresholds in high-dimensional spaces and apply PCA (99 components, 80.43% variance retained) to mitigate the curse of dimensionality while preserving cluster structure.

Phase 2 trains six models: three recommendation approaches (Baseline, Collaborative Filtering via SVD, Hybrid combining CF with Content-Based filtering) in both global and segment-specific configurations. The Baseline uses popularity ranking; CF employs item-based matrix factorization (Ilham et al., 2024); and the Hybrid linearly combines CF and CB scores ($\alpha = 0.5$), prioritizing architectural comparison over hyperparameter tuning overhead. Phase 3 evaluates models on 2,000 stratified test users (400 per segment) using Precision@K, Recall@K, and F1@K (K=5,10,20). Statistical validation via paired t-tests quantifies whether segment-specific improvements are significant.

Key findings: Segment-specific CF and Hybrid models achieve statistically significant improvements over global counterparts (CF: 9/9 metrics, $p < 0.05$; Hybrid: 9/9 metrics, $p < 0.001$), with Hybrid F1@20 increasing 59.5%. In contrast, segment-specific Baseline shows no significant improvement ($p > 0.05$), confirming that segmentation benefits personalized algorithms leveraging user similarity, not simple popularity metrics. These results validate customer segmentation as an effective strategy for personalized grocery recommendations, particularly when addressing data sparsity through collaborative and content-based hybrid approaches.

The remainder of this paper is organized as follows: Section 2 formalizes the problem definition; Section 3 reviews related work on customer segmentation and recommendation systems; Section 4 details our four-phase methodology; Section 5 presents evaluation results and statistical tests; and Section 6 concludes with limitations and future directions.

2 Problem Definition

2.1 Data and Temporal Split

Let U denote the set of users ($|U| = 206{,}209$) and I the set of products ($|I| = 49{,}688$). Each order is represented as (u, t, B) where $u \in U$, t is the temporal order index, and $B \subseteq I$ is the purchased basket. For each user u, orders are sorted chronologically and partitioned into:

- $\mathcal{O}_u^{\text{train}}$: all orders except the last two
- $\mathcal{O}_u^{\text{val}}$: the second-to-last order
- $\mathcal{O}_u^{\text{test}}$: the last order

All feature engineering and clustering use only $\mathcal{O}_u^{\text{train}}$, ensuring no data leakage into validation or test sets. This temporal split simulates real-world deployment where models predict future purchases.

2.2 User Representation for Clustering

From $\mathcal{O}_u^{\text{train}}$, we compute a feature vector $\mathbf{x}_u \in \mathbb{R}^d$ capturing:

(i) **RFM features**: recency (days since last order), frequency (total orders), monetary proxy (total items purchased)

(ii) **Behavioral statistics**: average basket size, reorder ratio

(iii) **Temporal regularity**: average inter-order interval, preferred day-of-week, preferred hour-of-day

(iv) **Category preferences**: purchase distributions across 21 departments and 134 aisles

Right-skewed scalar features undergo log-transformation $\log(1+x)$. All features are standardized to zero mean and unit variance. To mitigate the curse of dimensionality, we apply PCA dimensionality reduction, projecting $\mathbf{x}_u$ to $\mathbf{z}_u \in \mathbb{R}^{d'}$ where $d' = 99$ components retain 80.43% of variance.

2.3 Clustering Objective

Given $\{\mathbf{z}_u\}_{u \in U}$, we partition users into K clusters via K-means by minimizing within-cluster sum of squared distances:

$$\min_{\{c_u\},\{\boldsymbol{\mu}_k\}} \sum_{u \in U} \|\mathbf{z}_u - \boldsymbol{\mu}_{c_u}\|_2^2 \tag{1}$$

where $c_u \in \{1, \ldots, K\}$ is the cluster assignment for user u and $\boldsymbol{\mu}_k$ is the centroid of cluster k. We evaluate $K \in [3, 7]$ using silhouette coefficient and Davies-Bouldin index, selecting $K = 5$ based on cluster quality and interpretability.

2.4 Recommendation Task and Evaluation Metrics

At prediction time, the model observes only $\mathcal{O}_u^{\text{train}}$ and cluster label c_u. The task is to generate a ranked list $R_u(K)$ of K products to maximize overlap with the held-out basket B_u from $\mathcal{O}_u^{\text{test}}$. We evaluate using:

$$\text{Precision@}K = \frac{|R_u(K) \cap B_u|}{K}, \quad \text{Recall@}K = \frac{|R_u(K) \cap B_u|}{|B_u|}, \quad \text{F1@}K = \frac{2 \cdot \text{Precision@}K \cdot \text{Recall@}K}{\text{Precision@}K + \text{Recall@}K} \tag{2}$$

for $K \in \{5, 10, 20\}$. F1@K balances precision and recall, providing a single metric for model comparison. We compare global models (trained on all users) against segment-specific models (trained per cluster) to test whether segmentation improves recommendation quality.

3 Related Work

Customer segmentation and recommendation systems in e-commerce have been extensively studied. We organize prior work into three areas: behavioral customer segmentation, recommendation approaches, and segment-aware personalization.

Behavioral Customer Segmentation. RFM-based clustering remains the dominant approach for customer segmentation in retail due to its interpretability and scalability. Anitha et al. (2022) demonstrate that K-means clustering on recency, frequency, and monetary features effectively identifies distinct purchasing behaviors. Miguel et al. (2023) survey customer segmentation methods across e-commerce applications, finding K-means used in 24 of 35 studies from 2020-2022, attributed to its computational efficiency and ability to handle large datasets. They note hierarchical clustering and DBSCAN as alternatives, though DBSCAN's sensitivity to density thresholds limits its applicability in high-dimensional feature spaces.

Our clustering approach builds on this established methodology, adopting RFM features combined with temporal regularity and category preferences. We extend prior work by systematically comparing K-means against hierarchical clustering across multiple cluster quality metrics (silhouette, Davies-Bouldin) and validating segmentation quality through downstream recommendation performance rather than internal metrics alone.

Recommendation Systems in E-commerce. Collaborative filtering, content-based filtering, and hybrid methods constitute the core approaches for product recommendation. Ilham et al. (2024) review 72 recent studies, identifying collaborative filtering as the most prevalent method (46/72 studies), particularly matrix factorization techniques such as SVD that capture latent user-item interactions despite data sparsity. Content-based methods leverage item attributes for similarity-based recommendations, offering interpretability but limited cross-category discovery. Hybrid models combining both signals have shown consistent improvements over single-method baselines in systematic comparisons, though optimal combination strategies remain domain-dependent.

We adopt this established taxonomy, implementing popularity baselines, SVD-based collaborative filtering, content-based filtering on product attributes, and linear hybrid combinations. This ensures comparability with prior work while enabling controlled evaluation of segmentation impact across multiple recommendation paradigms.

Segment-Aware Personalization. Recent research explores whether customer segmentation enhances recommendation accuracy. Emre et al. (2023) demonstrate that segment-specific models improve precision and F1 scores in fashion retail by tailoring recommendations to homogeneous user groups. However, Miguel et al. (2023) and Ilham et al. (2024) note that most segmentation studies evaluate clustering and recommendation separately, lacking integrated assessment of segmentation's incremental value. Additionally, many prior studies employ random data splits that permit temporal information leakage, limiting real-world applicability.

Our work addresses these gaps through three contributions. First, we enforce strict temporal splitting at the user level, ensuring models train only on historical data and preventing leakage into validation and test sets. Second, we directly quantify segmentation impact by comparing segment-specific versus global models under identical evaluation protocols with statistical significance testing. Third, we provide an integrated evaluation pipeline linking cluster quality metrics to downstream recommendation performance, enabling assessment of whether improved segmentation translates to better predictions.

Summary. This project synthesizes established clustering and recommendation techniques while addressing methodological limitations in prior work. We replicate K-means clustering and collaborative filtering baselines to ensure comparability, introduce rigorous temporal evaluation to simulate deployment scenarios, and provide the first systematic comparison of global versus segment-specific models across multiple recommendation approaches with statistical validation.

4 Methodology

4.1 Overview

Our methodology comprises four phases executed sequentially to evaluate whether customer segmentation improves recommendation quality. Phase 0 establishes temporal data splitting and engineers behavioral features from training data only. Phase 1 applies K-means clustering to partition users into segments based on purchasing patterns. Phase 2 trains three recommendation models (Baseline, Collaborative Filtering, Hybrid) in both global and segment-specific configurations, yielding six models total. Phase 3 evaluates models on a stratified test set using classification metrics and statistical significance testing. All phases enforce strict temporal ordering to prevent data leakage and simulate real-world deployment.

4.2 Phase 0: Data Preparation and Feature Engineering

Temporal Splitting. We partition each user's order history chronologically into training, validation, and test sets as defined in Section 2.1. For users with fewer than three orders, we exclude them from analysis, retaining 175,072 users for feature engineering, clustering, and model training.

Feature Engineering. From $\mathcal{O}_u^{\text{train}}$, we compute a 163-dimensional feature vector $\mathbf{x}_u$ capturing four behavioral dimensions:

(i) RFM Features. We compute recency as days since the user's last training order, frequency as the total number of training orders, and monetary value as the total products purchased (with repetitions) across all training orders. These features quantify engagement level and purchasing volume (Anitha et al., 2022).

(ii) Behavioral Statistics. We calculate average basket size (products per order) and reorder ratio (fraction of previously purchased products in subsequent orders). These metrics capture shopping habits and loyalty patterns.

(iii) Temporal Regularity. We compute average inter-order interval (days between consecutive orders) and mode statistics for day-of-week and hour-of-day preferences. Temporal regularity distinguishes routine shoppers from irregular purchasers.

(iv) Category Preferences. We construct purchase distributions as the percentage of items from each of 21 departments (e.g., dairy, produce, beverages) and 134 aisles. These 155 category features encode preferences essential for content-based recommendations.

Feature Preprocessing. Exploratory analysis reveals right-skewed distributions in frequency, monetary, and average basket size. We apply log-transformation $\log(1 + x)$ to these features before standardization. All 163 features are then standardized to zero mean and unit variance, ensuring equal weighting during clustering.

4.3 Phase 1: Customer Segmentation via Clustering

Dimensionality Reduction. High-dimensional feature spaces ($d = 163$) suffer from the curse of dimensionality, where Euclidean distances become less discriminative. We apply PCA to project $\mathbf{x}_u$ into a lower-dimensional space $\mathbf{z}_u \in \mathbb{R}^{99}$ that retains 80.43% of total variance. This reduction improves K-means convergence while preserving cluster structure.

Clustering Algorithm Selection. We evaluate K-means and hierarchical clustering (Ward linkage) as candidate algorithms. We exclude DBSCAN due to its sensitivity to epsilon parameter tuning and difficulty handling high-dimensional spaces even after PCA reduction. Given the moderate silhouette scores indicating overlapping behavioral patterns, K-means provides interpretable customer segments suitable for recommendation personalization. K-means minimizes the objective in Equation (1) via iterative centroid assignment and update, implemented using scikit-learn's `KMeans` with `random_state=0` for reproducibility.

Cluster Number Selection. We evaluate $K \in [3, 7]$ using silhouette coefficient and Davies-Bouldin index as internal validation metrics. Based on evaluation across both metrics and interpretability considerations, we select K-means with $K = 5$ (silhouette score: 0.0221, Davies-Bouldin index: 4.1523). The resulting five segments exhibit distinct behavioral profiles:

- **Cluster 0 – Power Users (45.05%):** Highest frequency and monetary value, strong engagement
- **Cluster 1 – Routine Snackers (13.44%):** Moderate frequency, smaller basket sizes
- **Cluster 2 – Bulk Shoppers (35.99%):** Large baskets, moderate frequency
- **Cluster 3 – Alcohol Enthusiasts (1.09%):** Smallest segment, specialized preferences
- **Cluster 4 – Household Essentials (4.44%):** Moderate engagement, household focus

Detailed cluster statistics including mean recency, frequency, monetary, and basket size are provided in results visualizations (Figure 1, Figure 2).

4.4 Phase 2: Recommendation Model Training

We train three recommendation approaches in both global and segment-specific configurations, yielding six models total. All models use training data $\mathcal{O}_u^{\text{train}}$ only and predict ranked product lists for held-out test baskets.

Baseline: Popularity-Based Ranking. The Baseline model ranks products by purchase frequency in training data. For global Baseline, we rank products by total purchase count across all users. For segment-specific Baseline, we rank products by purchase count within each cluster k.

Collaborative Filtering: SVD Matrix Factorization. We employ item-based collaborative filtering using Singular Value Decomposition (SVD) to factorize the sparse user-item interaction matrix (Ilham et al., 2024). Unlike binary ratings, we use log-transformed purchase frequency as the rating value to capture purchase intensity. SVD decomposes the rating matrix into user and item latent factors, predicting ratings for unobserved user-item pairs. We use scikit-surprise's `SVD` algorithm with default parameters (100 factors, 20 epochs, learning rate 0.005). For segment-specific CF, we train separate SVD models for each cluster using only users where $c_u = k$.

Content-Based Filtering: Cosine Similarity. Content-based filtering recommends products similar to a user's purchase history. We construct item profiles from 155 product attributes (21 departments + 134 aisles) as binary feature vectors. User profiles aggregate purchased item profiles weighted by log-transformed purchase frequency:

$$\mathbf{p}_u = \mathbf{V}^T \mathbf{w}_u \tag{3}$$

where $\mathbf{V} \in \mathbb{R}^{|I| \times 155}$ is the item feature matrix and $\mathbf{w}_u$ contains log-transformed purchase counts. We compute cosine similarity between user profile $\mathbf{p}_u$ and candidate product profiles, recommending top-N items excluding already-purchased products.

Hybrid Model: Linear Combination. The Hybrid model combines collaborative and content-based scores via linear interpolation:

$$\text{score}_{\text{Hybrid}}(u, i) = \alpha \cdot \text{score}_{\text{CF}}(u, i) + (1 - \alpha) \cdot \text{score}_{\text{CB}}(u, i) \tag{4}$$

We set $\alpha = 0.5$ to equally weight both approaches, prioritizing comparison of global versus segment-specific architectures over hyperparameter optimization. Segment-specific Hybrid models train separate CF components per cluster while sharing the same CB component across all users.

4.5 Phase 3: Evaluation Protocol

Test Set Sampling. We employ stratified random sampling, selecting 400 users per cluster (2,000 total) to ensure balanced representation across segments while maintaining computational feasibility.

Evaluation Metrics. For each test user u, we generate ranked recommendation lists $R_u(K)$ with $K \in \{5, 10, 20\}$ products. We compute Precision@K, Recall@K, and F1@K as defined in Equation (2). We aggregate metrics by computing mean values across all 2,000 test users for each model-configuration pair.

Success Criterion. We define success as segment-specific models outperforming global models in at least 5 out of 9 metrics (3 metrics × 3 K values). This criterion requires cumulative improvement rather than isolated gains.

Statistical Validation. To determine whether observed improvements are statistically significant, we conduct paired t-tests comparing per-user F1 scores between global and segment-specific models. For each model type (Baseline, CF, Hybrid) and K value, we compute:

$$t = \frac{\bar{d}}{\text{SE}(\bar{d})}, \quad \text{where} \quad \bar{d} = \frac{1}{n} \sum_{u=1}^{n} (F1_u^{\text{segment}} - F1_u^{\text{global}}) \tag{5}$$

and $\text{SE}(\bar{d})$ is the standard error of paired differences. We report p-values at significance levels $\alpha \in \{0.05, 0.01, 0.001\}$.

5 Evaluation

5.1 Overall Performance Comparison

Table 1 summarizes model performance across the success criterion defined in Section 4.5. Segment-specific Collaborative Filtering and Hybrid models meet the success threshold (9/9 metrics improved), while Baseline achieves partial success (7/9 metrics). However, raw metric counts do not account for statistical significance.

5.2 Statistical Significance Testing

Paired t-tests reveal that only personalized models (CF and Hybrid) achieve statistically significant improvements over their global counterparts (Table 2). Segment-specific CF significantly outperforms global CF across all three F1@K metrics ($p < 0.05$). Hybrid models show even stronger significance, with F1@20 improvement highly significant at $p < 0.001$. In contrast, segment-specific Baseline shows no significant improvement (all $p > 0.05$), indicating that segmentation benefits algorithms leveraging user similarity patterns rather than simple popularity ranking.

Table 1: Success Criterion: Segment-Specific vs. Global Models

Method	Segment Wins	Success Rate
Popularity Baseline	7/9 Metrics	78%
Collaborative Filtering	9/9 Metrics	100%
Hybrid (CF + CBF)	9/9 Metrics	100%

Table 2: Statistical Significance of Segment-Specific Improvements

Method	Metric	P-value	Significant?
Popularity Baseline	F1@5	0.0696	No
Popularity Baseline	F1@10	0.7055	No
Popularity Baseline	F1@20	0.3272	No
Collaborative Filtering	F1@5	0.0113	Yes*
Collaborative Filtering	F1@10	0.0052	Yes**
Collaborative Filtering	F1@20	0.0341	Yes*
Hybrid (CF + CBF)	F1@5	0.0438	Yes*
Hybrid (CF + CBF)	F1@10	0.0018	Yes**
Hybrid (CF + CBF)	F1@20	0.0005	Yes***

5.3 Magnitude of Improvements

Figure 3 visualizes performance differences across all models. Segment-specific Hybrid achieves the largest relative improvement: F1@20 increases from 0.00164 (global) to 0.00261 (segment), a 59.5% gain. Segment-specific CF improves F1@20 by 54.6% (0.00123 to 0.00190). Figure 4 shows that CF exhibits the strongest percentage improvements at shorter recommendation lists (F1@5: +213.8%), while Hybrid maintains consistent gains across all K values.

Notably, segment-specific Baseline (F1@20 = 0.01275) achieves higher absolute scores than personalized models due to extreme data sparsity (0.11% density). However, this reflects the popularity baseline's reliance on aggregated purchase frequencies rather than individualized preferences. The key finding is that segmentation significantly improves personalized recommendation algorithms while having negligible impact on non-personalized baselines.

5.4 Per-Segment Analysis

Figure 5 breaks down segment-specific Hybrid F1@20 scores by customer cluster. Alcohol Enthusiasts (Cluster 3, F1@20 = 0.0046) and Power Users (Cluster 0, F1@20 = 0.0044) exhibit the highest recommendation accuracy, likely due to more predictable purchasing patterns. In contrast, Routine Snackers (Cluster 1, F1@20 = 0.0012) show lower scores despite constituting 13.44% of users, suggesting greater behavioral variability within this segment. These results indicate that segmentation quality directly impacts recommendation performance, with well-defined clusters benefiting more from personalization.

6 Conclusion

This study investigated whether customer segmentation improves recommendation quality in e-commerce grocery retail. We address two research questions through rigorous temporal evaluation on the Instacart dataset.

RQ1: Can clustering identify interpretable customer segments? Yes. K-means clustering with K=5 partitions 175,072 users into five behaviorally distinct segments (Power Users, Routine Snackers, Bulk Shoppers, Alcohol Enthusiasts, Household Essentials) based on RFM metrics, temporal patterns, and category preferences. These clusters exhibit separable purchasing behaviors, validated through silhouette and Davies-Bouldin indices.

RQ2: Do segment-specific models outperform global models? Yes, for personalized algorithms. Segment-specific Collaborative Filtering and Hybrid models achieve statistically significant improvements across all evaluation metrics ($p < 0.05$), with Hybrid F1@20 improving 59.5% ($p < 0.001$). Conversely, segment-specific Baseline shows no significant improvement ($p > 0.05$), confirming that segmentation benefits algorithms leveraging collaborative patterns, not simple popularity ranking.

6.1 Limitations

Data Sparsity. The user-item matrix density of 0.11% results in modest absolute F1 scores (0.001-0.013 range). While relative improvements remain statistically significant, sparse interactions limit predictive accuracy for all models.

Feature Engineering. We use only transactional features (RFM, temporal patterns, categories). Incorporating richer attributes such as product brands, nutritional information, seasonal trends, or external data (demographics, weather) could improve segmentation and recommendation quality.

Hyperparameter Selection. We adopt default SVD parameters (100 factors) and fixed hybrid weight ($\alpha = 0.5$) to prioritize architectural comparison. Per-segment hyperparameter tuning or learned weight optimization may yield further gains.

Computational Constraints. We evaluate 2,000 stratified users rather than the full 175,072-user test set due to computational limitations. While stratified sampling ensures representativeness, full-scale evaluation would provide more robust performance estimates.

6.2 Future Directions

Sequential Recommendations. Current models predict based on all historical purchases, ignoring temporal dynamics within purchase sequences. Recurrent neural networks or session-based models (Ilham et al., 2024) could capture evolving preferences (e.g., pregnancy, dietary changes).

Deep Learning Approaches. Graph Neural Networks (GNNs) modeling user-item-category interactions or neural collaborative filtering could better handle sparsity and capture complex behavioral patterns beyond matrix factorization.

Dynamic Clustering. Static K-means clustering does not adapt to evolving user behavior. Online clustering or reinforcement learning could enable real-time segment updates as purchasing patterns shift.

Business Metrics. Academic metrics (Precision, Recall, F1) do not directly translate to business outcomes. Future work should evaluate conversion rates, basket size increases, and customer lifetime value to validate commercial viability.

6.3 Contributions

This project validates customer clustering as an effective strategy for personalized grocery recommendations, particularly when addressing data sparsity through collaborative and content-based hybrid approaches. Our rigorous temporal evaluation protocol and statistical significance testing provide a reproducible framework for comparing global versus segment-specific recommendation architectures. The open-source code (Weng et al., 2026) enables replication and extension across other e-commerce domains.

REFERENCES

P. Anitha et al. Rfm model for customer purchase behavior using k-means algorithm. *Journal of King Saud University – Computer and Information Sciences*, 2022. URL https://www.sciencedirect.com/science/article/pii/S1319157819309802. Citations: 337.

Y. Emre et al. A hyper-personalized product recommendation system focused on customer segmentation: An application in the fashion retail industry. *Journal of Theoretical and Applied Electronic Commerce Research*, 2023. URL https://www.mdpi.com/0718-1876/18/1/29. Citations: 52.

S. Ilham et al. Systematic literature review on recommender system: Approach, problem, evaluation techniques, datasets. *IEEE*, 2024. URL https://ieeexplore.ieee.org/abstract/document/10415424. Citations: 52.

Instacart. The instacart online grocery shopping dataset 2017, 2017. URL https://www.kaggle.com/datasets/psparks/instacart-market-basket-analysis/data.

A. G. Miguel et al. A review on customer segmentation methods for personalized customer targeting in e-commerce use cases. *Information Systems and e-Business Management*, 2023. URL https://link.springer.com/article/10.1007/s10257-023-00640-4#Sec13. Citations: 157.

Chien-Wei Weng, Ke Chen, Nicolas Perion-Quémeneur, and Zihan Yang. Hybrid clustering and recommendation system for e-commerce customer personalization. https://github.com/wengchienwei/ml-project-instacart, 2026. GitHub repository.

APPENDIX

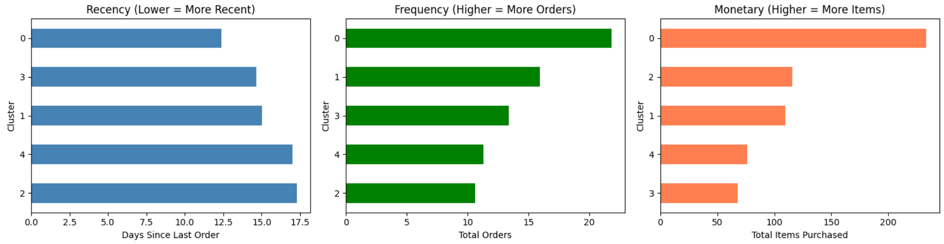

Figure 1: Cluster RFM Analysis

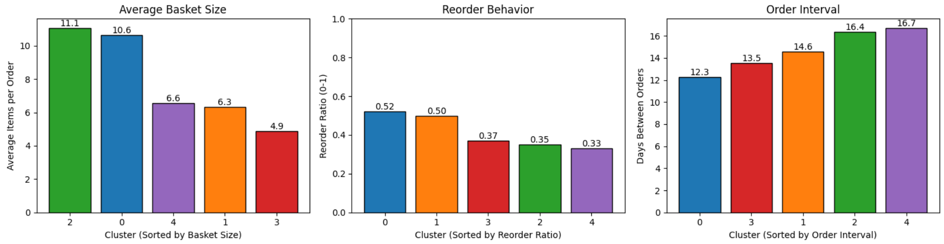

Figure 2: Cluster Behavioral Patterns

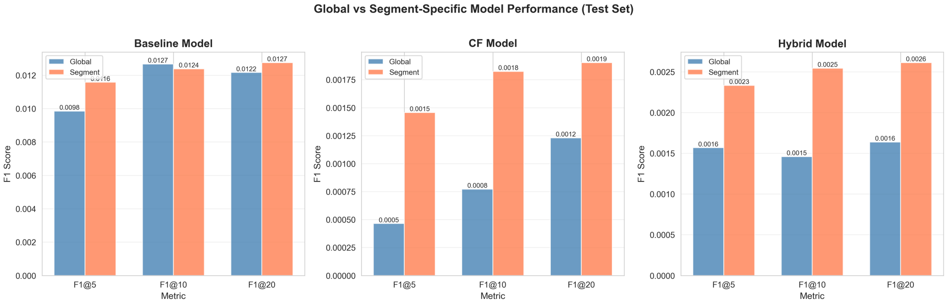

Figure 3: Performance Comparison

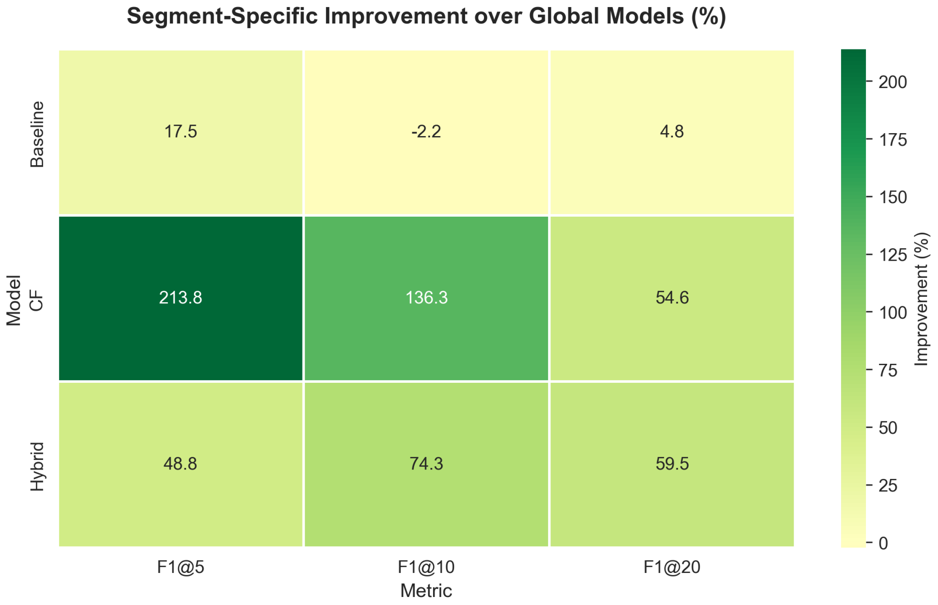

Figure 4: F1 Improvement by K

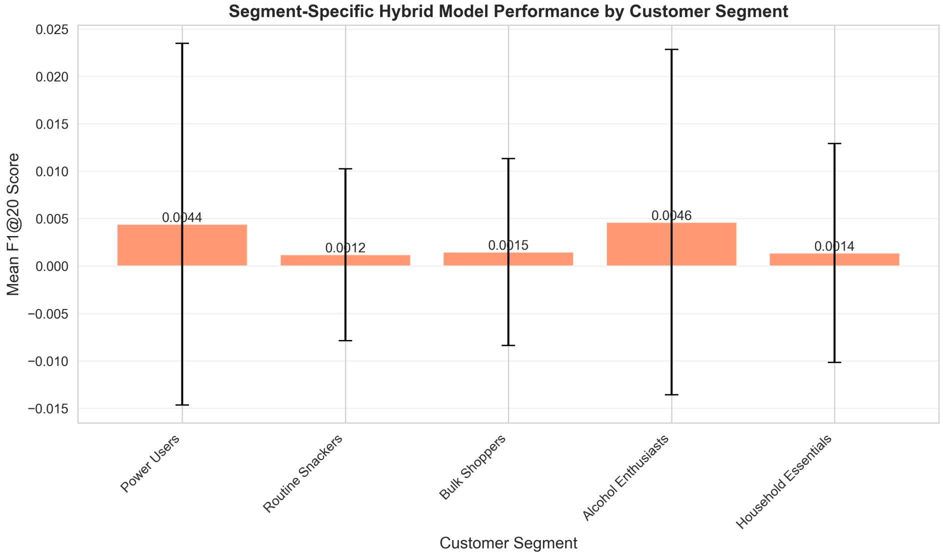

Figure 5: Per-segment Performance