GROCERY STORE RESEARCH

The Price Pulse of Long Island: Final Implementation, Modeling & Evaluation

Predicting 7-Day Grocery Catalog Price Changes (>=5%) Across Longitudinal Retail Snapshots

Author & Project Owner: Frank Valdez | Archive Window: May 16 – September 6, 2026 | Phase 3 Final Submission

1,031,264

Raw Catalog Rows Across 82 Snapshots

254,453

Exact-7-Day Labeled King Kullen Pairs

0.793 / 0.543

Best Test ROC-AUC / PR-AUC (Tree Ensembles)

0.478 / 0.715

Best Test F1 / Balanced Accuracy (HistGB)

1. Motivation, Problem Statement & Research Objective

Why longitudinal grocery catalog analysis matters and how the supervised prediction task is formulated

Domain Motivation & Real-World Significance

- Supermarkets use High-Low ('Hi-Lo') promotional pricing: temporary weekly sales followed by sharp rollbacks to regular prices.
- Weekly 10%–50% swings on staples directly impact household budgets.
- Public online catalogs expose price and category data across chains, but lack explicit promotion expiration dates.
- Predicting next-week price shifts powers consumer price alerts and automated catalog auditing.

Exact Supervised Research Question & Target Formula

- Research Question: Can past-only product attributes, current discount depth, and price history predict a >=5% price change exactly 7 days later?
- Analytical Unit: Retailer source, string UPC/product key, and snapshot date t.
- Exact Integer-Cent Label: y = 1 iff 100 * |future_cents(t+7) - current_cents(t)| >= 5 * current_cents(t).
- Unobserved future listings remain unlabeled—never zero-filled as 'no change'.

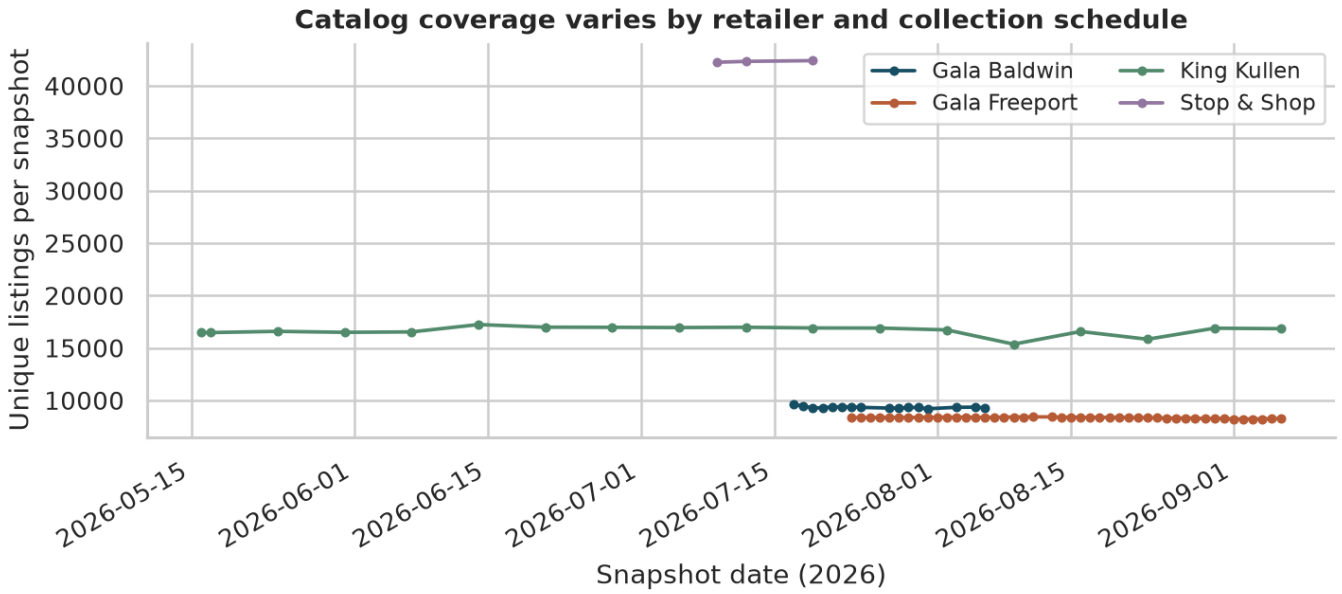
End-to-End Phase 3 Research Pipeline

- 1. Multi-Retailer Consolidation: 82 raw JSONL/gzipped snapshots (161.58 MB) across King Kullen, Stop & Shop, Gala Baldwin, and Gala Freeport -> 952,034 unique observations.
- 2. Leak-Free Feature Engineering: 10 past-only numeric features (price, discount depth, 7d return, 4-row mean |return| & volatility, coverage, gaps) + 4 missingness indicators + 158 one-hot categories = 172 columns.
- 3. Expanding-Window Temporal Validation & Held-Out Test Evaluation: Tune 5 model families across 5 expanding training folds (207,942 train rows through 2026-08-16) and evaluate on 3 held-out test weeks (46,511 test rows, 2026-08-23 to 2026-09-06).
- 4. Granular Subgroup & Unsupervised Regime Analysis: Evaluate stability across weeks, categories, cold-start UPCs, and 4 K-Means product pricing regimes.

2. Data Overview: Four Long Island Grocery Catalog Archives

1,031,264 raw records across 82 snapshots (May 16 – September 6, 2026) consolidated into 952,034 unique observations

Retailer Archive	2026 Coverage	Snapshots	Raw Rows	Repeated Rows	Clean Rows	Unique IDs	Stored
King Kullen (Primary)	2026-05-16 to 2026-09-06	18	378,987	79,230 (20.9%)	299,757	21,040	98.82 MB
Stop & Shop	2026-07-09 to 2026-07-19	3	127,002	0 (0.0%)	127,002	43,216	17.98 MB
Gala Baldwin	2026-07-17 to 2026-08-06	16	149,505	0 (0.0%)	149,505	11,822	10.76 MB
Gala Freeport	2026-07-23 to 2026-09-06	45	375,770	0 (0.0%)	375,770	9,797	34.01 MB
Total / Consolidated	May 16 – Sep 6, 2026	82	1,031,264	79,230	952,034	85,875	161.58 MB

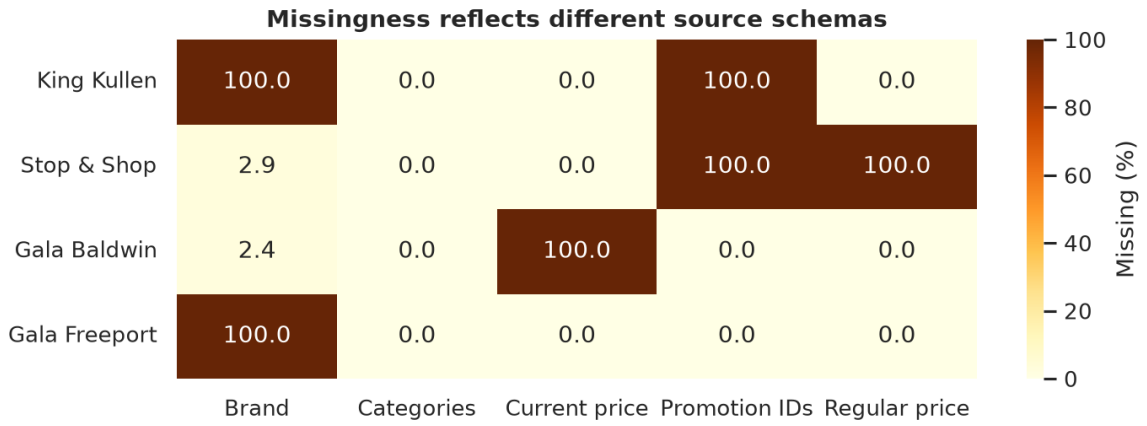


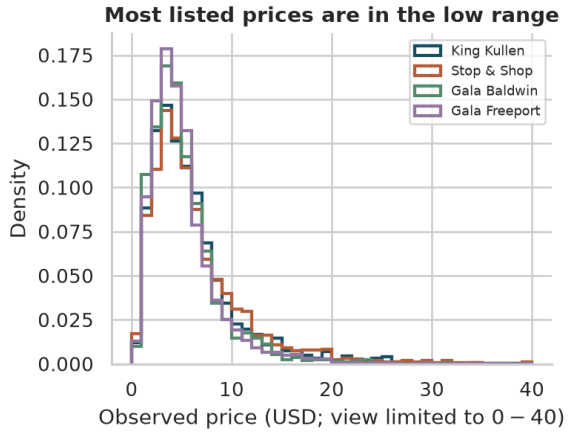
Why King Kullen Is the Supervised Benchmark

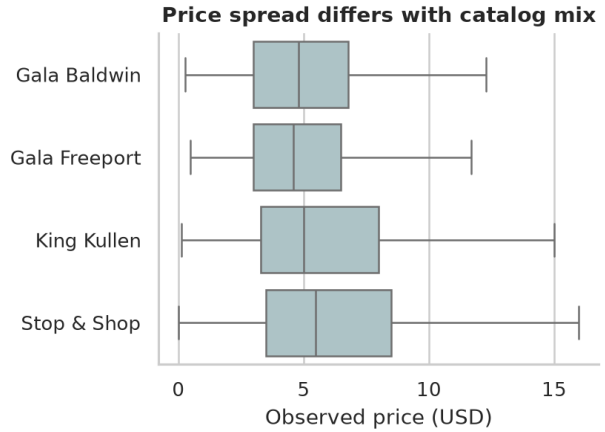
- Weekly Cadence: 18 snapshots spanning 16 exact-7-day target intervals from May to September 2026.
- Complete Dual Pricing: Every observation provides both current_price and regular_price, enabling exact discount_depth.
- Cross-Retailer Discipline: Stop & Shop (3 dates) and Gala Baldwin (regular price only) support descriptive EDA without unverifiable cross-store entity matching.

3. Data Cleaning, Schema Missingness & Outlier Handling

Reconciling multi-category duplicates, timezone boundaries, sentinel fields, and right-skewed price tails







Key Cleaning & Quality Audit Results

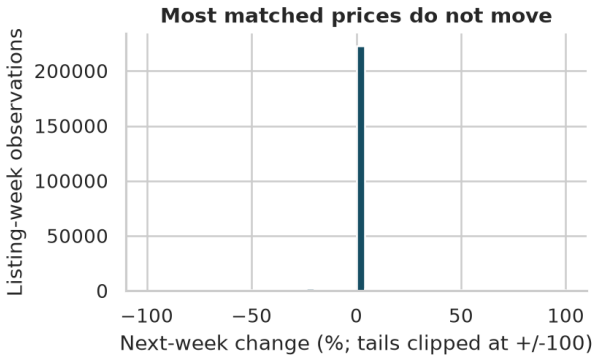
- Multi-Category Deduplication: 79,230 repeated King Kullen rows had 0 price conflicts; consolidated by unioning category sets.
- Timezone Reconciliation: 51,861 evening rows crossing UTC midnight matched America/New_York archive dates and were retained.
- Sentinel Removal: King Kullen department is 100% 'UNAVAILABLE' -> excluded; alphabetical primary_category used.
- Above-Regular Flag: 13,763 Freeport rows with current > regular price flagged, never swapped.

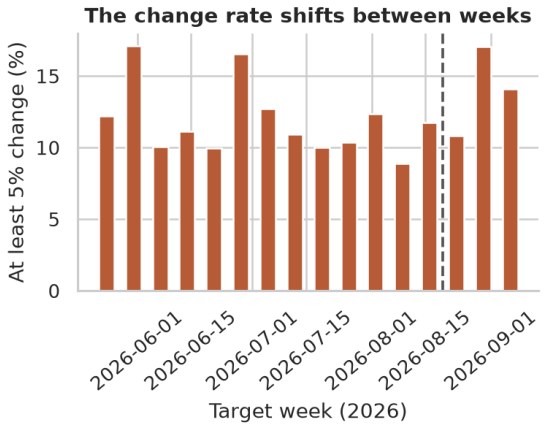
Right-Skewed Distributions & Outlier Policy

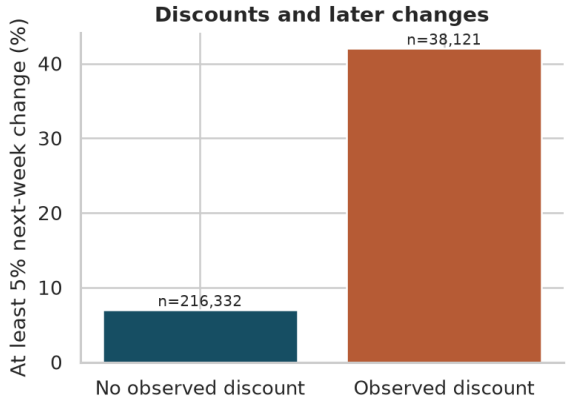
- Latest Medians: King Kullen $4.99 (IQR $3.29–$7.99, max $249.00); Stop & Shop $5.49; Gala Baldwin $4.79; Gala Freeport $4.59.
- Upper IQR Fence: $15.04 in King Kullen (915 listings above fence).
- Domain Preservation: High-priced items are genuine catering platters ($249 Turkey Dinner) and bulk packs—retained without winsorizing.

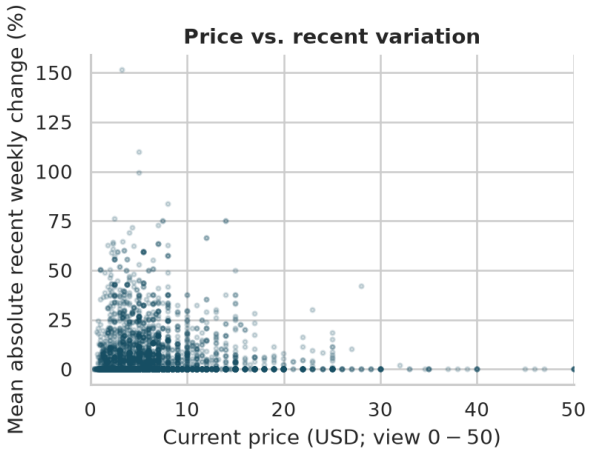
4. Exploratory Data Analysis: Weekly Prevalence & Discount Dynamics

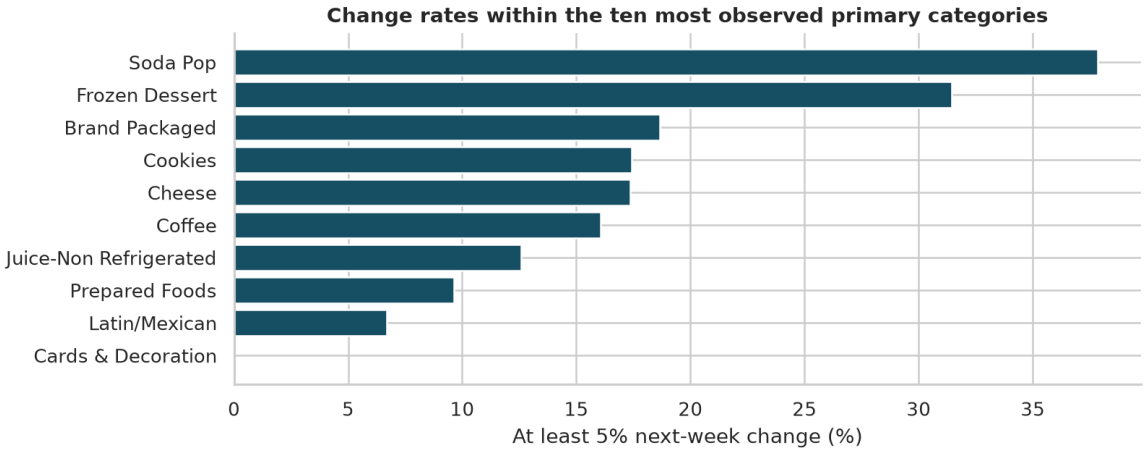
Promotional discounts drive a 6x increase in next-week price movement (42.04% vs. 7.00%)











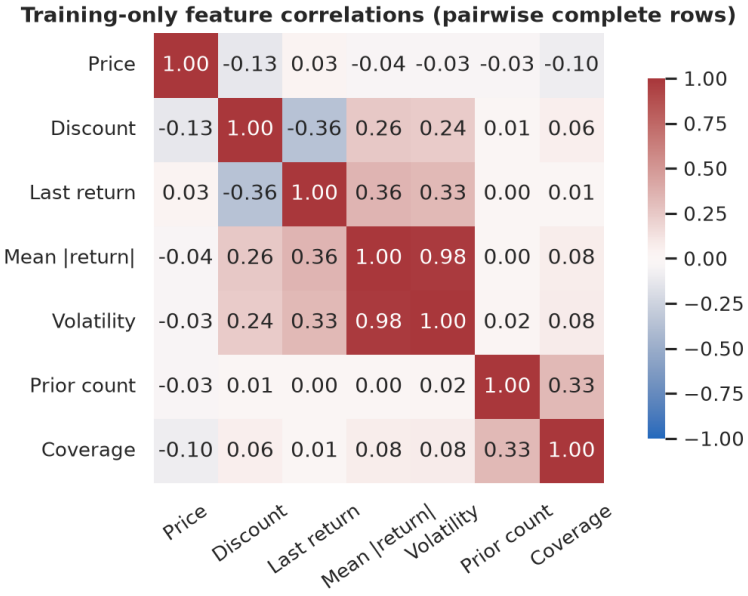
Core Exploratory Insights Shaping Phase 3 Modeling

- Price Stickiness: 87.52% (222,695) of matched King Kullen pairs have 0% change; 12.25% (31,176) change by >=5%.
- Weekly Regime Shifts: Positive rate fluctuates between 8.91% (Aug 9) and 17.12% (May 31) / 17.06% (Aug 30).
- Discount Rollback Effect: Currently discounted items change >=5% next week 42.04% of the time vs. 7.00% when undiscounted.
- Category Spread: Soda Pop (37.85%) and Frozen Dessert (31.46%) cycle rapidly; Cards & Decoration (0.00%) remain static.

5. Leak-Free Feature Engineering & Chronological Preprocessing

10 past-only numeric predictors, 4 missingness indicators, and 158 encoded categories (172 total columns)

Feature Group	Engineered Predictors	Train Miss	Test Miss	Transformation Pipeline
Current Price & Promo	current_price, discount_depth	0.00%	0.00%	Median impute + StandardScaler (train mean $6.57, discount 3.26%)
Recent Return (7d)	return_7d	11.19%	6.96%	Train median impute + binary missingindicator_return_7d + scale
Rolling 4-Obs Magnitude	mean_abs_change_4	9.69%	1.30%	Train median impute + binary missingindicator_mean_abs_change_4 + scale
Rolling 4-Obs Volatility	volatility_4 (requires >=2 returns)	19.21%	2.32%	Train median impute + binary missingindicator_volatility_4 + scale
Observation Cadence	valid_changes_4, gap_days	0% / 1.85%	0% / 1.22%	Train median impute + binary missingindicator_gap_days + scale
Listing Maturity	prior_observations, history_age_days, coverage_to_date	0.00%	0.00%	Train median impute + StandardScaler (mean coverage 92.8%)
Product Category	primary_category (292 train levels)	0.00%	0.00%	OneHotEncoder(min_frequency=100, handle_unknown='infrequent_if_exist') -> 158 cols



Chronological Split & Imbalance Protocol

- Training Set (13 Target Weeks, May 24 – Aug 16): 207,942 rows, 24,685 positives (11.87% positive rate).
- Held-Out Test Set (3 Target Weeks, Aug 23 – Sep 6): 46,511 rows, 6,491 positives (13.96% positive rate).
- Strict Temporal Boundary: max(train.target_date) = 2026-08-16 <= min(test.date) = 2026-08-16. Zero future leakage.
- Class Imbalance Handling: Compare balanced weights (0: 0.567, 1: 4.212), unweighted training, and CV-tuned thresholding.

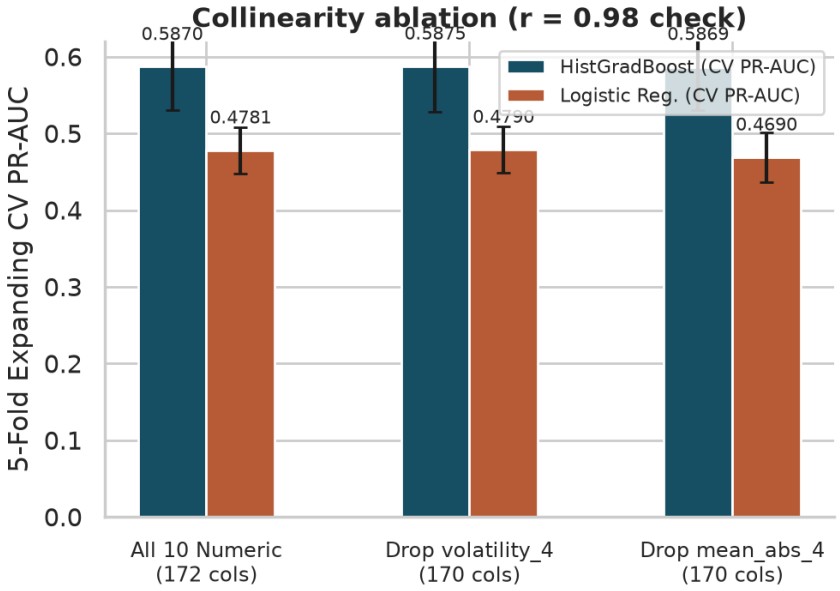
6. Data Modeling: Algorithms Introduced & Selection Justification

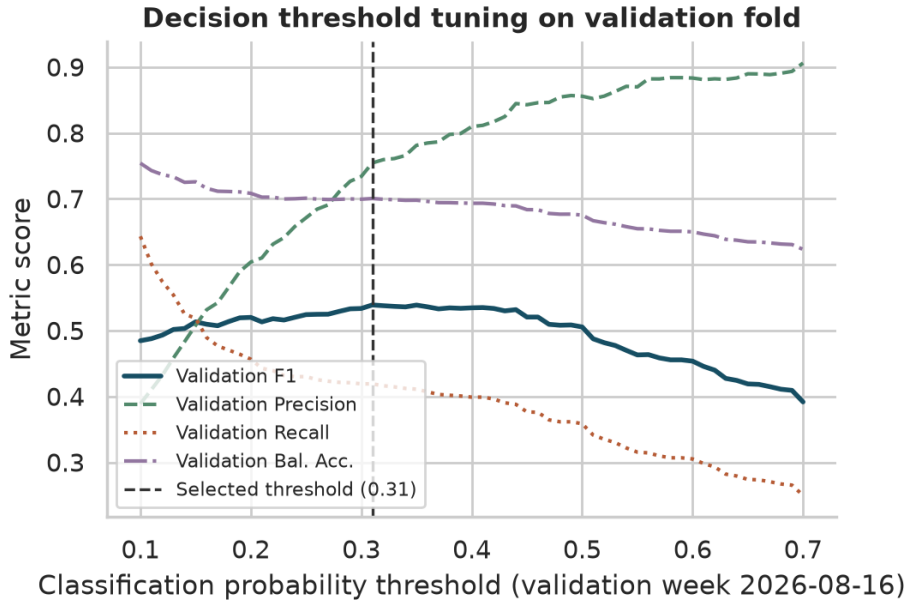
Comparing two domain baselines, five supervised classification families, and unsupervised K-Means clustering

Algorithm / Family	Formulation & Inductive Bias	Why Selected for Grocery Price Prediction
1. Domain Baselines	No-Change Majority (y=0) & Observed Discount Heuristic (y=1 iff discount_depth > 0)	Establishes whether ML models add predictive value beyond trivial majority voting or a simple 'on-sale today' rule.
2. Logistic Regression	L2-penalized log-odds linear classifier over 172 standardized & one-hot features	Provides calibrated baseline probabilities and interpretable directional log-odds coefficients for each feature/category.
3. Linear SVM (LinearSVC)	Maximum-margin hyperplane minimizing L2-regularized squared-hinge loss	Tests whether margin maximization improves linear separation in high-dimensional sparse one-hot category space.
4. Decision Tree	CART recursive axis-aligned partitioning minimizing weighted Gini impurity	Captures non-linear threshold rules (e.g., discount_depth > 0 AND return_7d < 0) with transparent decision paths.
5. Random Forest	Bagged ensemble of 200 deep trees with bootstrap sampling & feature subsampling	Reduces high variance of single decision trees while modeling complex interactions between category and price history.
6. HistGradientBoosting	LightGBM-style histogram-binned gradient boosted trees with L2 leaf regularization	State-of-the-art tabular learner; sequentially corrects residual errors and handles asymmetric promotional dynamics.
7. K-Means (Unsupervised)	Centroid clustering on 5 standardized training UPC pricing-behavior descriptors	Discovers latent product pricing regimes (staples vs. Hi-Lo promotional items) evaluated via Silhouette & Davies-Bouldin.

7. Expanding-Window Temporal CV, Collinearity Check & Tuning

Refitting preprocessing inside 5 expanding training folds prevents temporal leakage during model selection





Collinearity Ablation (mean_abs_change_4 vs. volatility_4, r = 0.98)

- All 10 Numeric (172 cols): HistGB CV PR-AUC = 0.5870, F1 = 0.4681; LogReg CV PR-AUC = 0.4781.
- Drop volatility_4 (170 cols): Slightly aids LogReg BalAcc (0.7507); HistGB F1 drops to 0.4633.
- Why Retain All 10: Their missingness indicators differ (9.69% vs. 19.21%), encoding whether an item has 1 vs. >=2 prior changes.

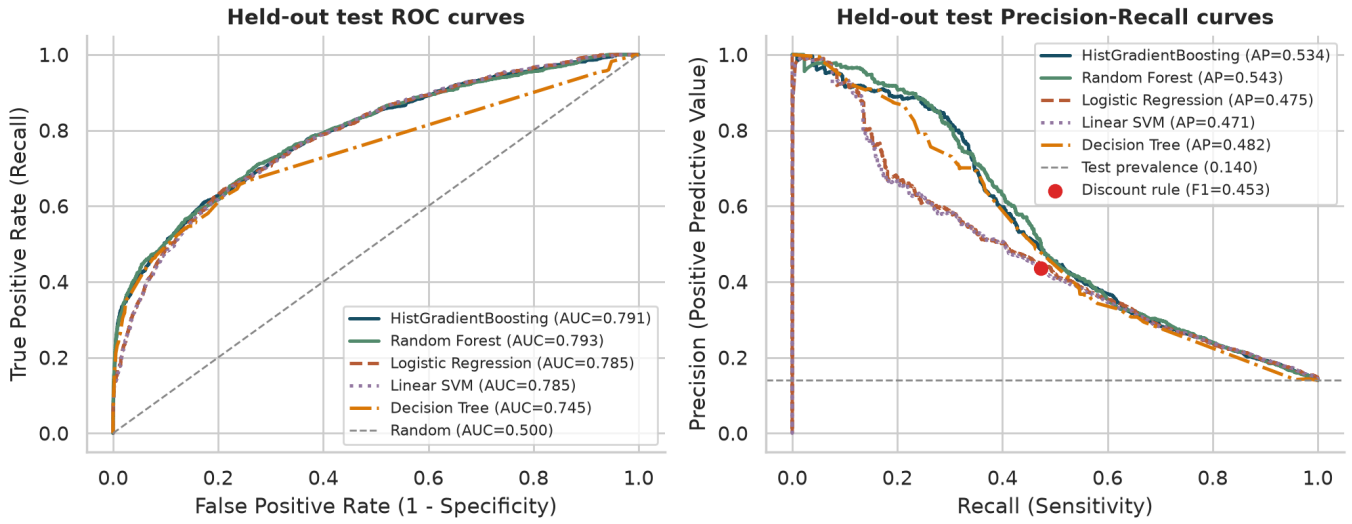
Hyperparameter & Threshold Selection (5 Expanding Folds)

- Tree Ensembles Lead CV: Random Forest (PR-AUC 0.5939 bal / 0.5957 unw) & HistGB (0.5879 bal / 0.5920 unw) >> Decision Tree (0.5326) >> LogReg (0.4785) & SVM (0.4762).
- Validation Threshold Optimization: tau* = 0.31 maximizes validation F1 on fold week 2026-08-16.

8. Held-Out Test Set Results & Performance Comparison

Evaluating all models on 46,511 held-out observations across 2026-08-23, 2026-08-30, and 2026-09-06 (13.96% positive)

Model / Configuration	Regime	Acc.	BalAcc	Prec.	Recall	Spec.	F1	ROC-AUC	PR-AUC
No-Change Baseline	Majority (y=0)	0.8604	0.5000	0.0000	0.0000	1.0000	0.0000	0.5000	0.1396
Observed Discount Heuristic	Domain Rule	0.8408	0.6868	0.4352	0.4731	0.9004	0.4534	0.6941	0.3503
Decision Tree (depth=8)	Balanced Weights	0.8029	0.6995	0.3648	0.5562	0.8429	0.4406	0.7451	0.4821
Linear SVM (C=0.01)	Balanced Weights	0.8164	0.7010	0.3871	0.5409	0.8611	0.4513	0.7845	0.4706
Logistic Regression (C=0.1)	Balanced Weights	0.8139	0.7035	0.3837	0.5505	0.8566	0.4522	0.7851	0.4746
Random Forest (200 trees)	Balanced Weights	0.8070	0.7120	0.3759	0.5802	0.8438	0.4562	0.7929	0.5428
HistGradientBoosting (leaves=31)	Balanced Weights	0.8057	0.7153	0.3753	0.5899	0.8408	0.4588	0.7912	0.5340
HistGradientBoosting (leaves=31)	Unweighted (0.50)	0.8891	0.6637	0.7066	0.3509	0.9764	0.4690	0.7935	0.5364
HistGradientBoosting (tau*=0.31)	CV Threshold	0.8786	0.6777	0.5975	0.3990	0.9564	0.4785	0.7935	0.5364

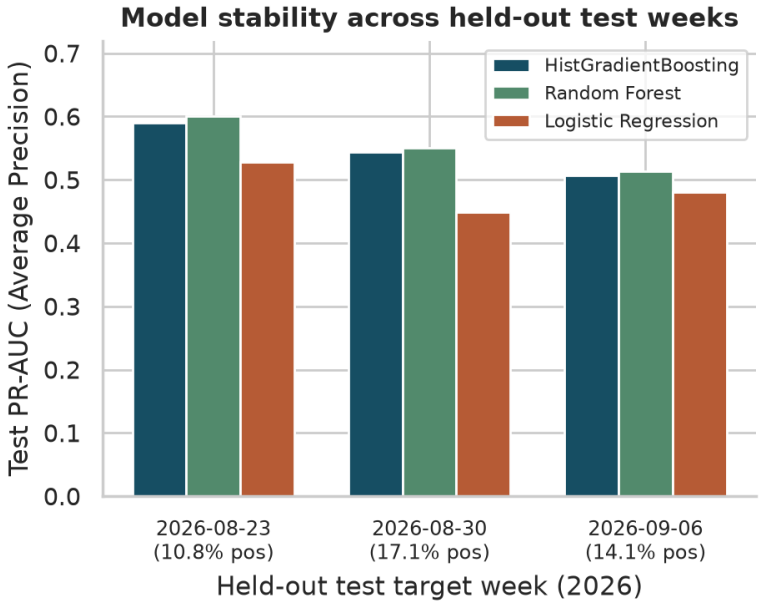


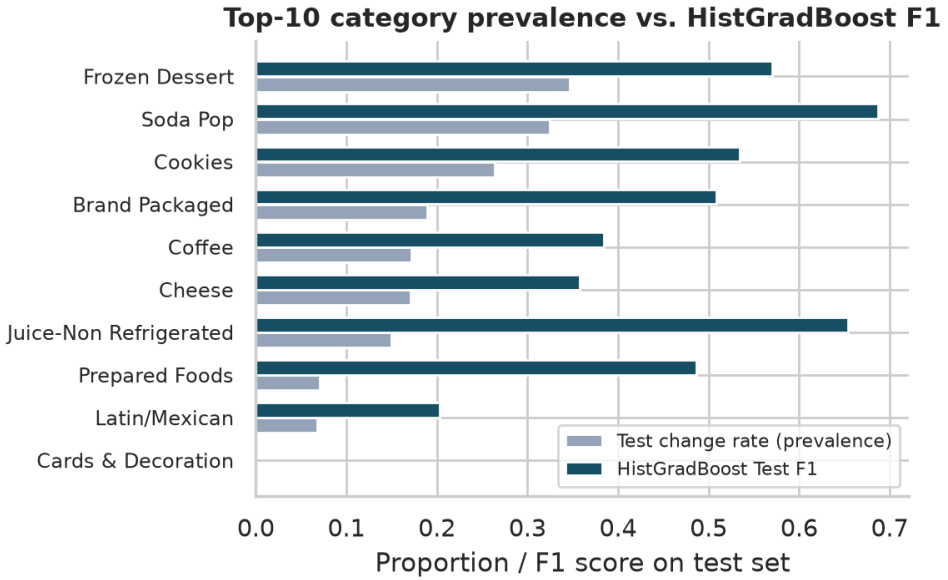
Key Test Set Findings

- Tree Ensembles Lead: Random Forest (PR-AUC 0.5428, ROC-AUC 0.7929) & HistGB (PR-AUC 0.5364, ROC-AUC 0.7935) beat linear models by +6 PR-AUC pts.
- Operating Regimes: HistGB (Balanced) achieves top Balanced Accuracy (0.7153); CV-Threshold (0.31) achieves top F1 (0.4785).
- Gain Over Heuristic: +54% relative PR-AUC lift over the Discount Rule (0.3503 -> 0.5364).

9. Subgroup Evaluation: Weekly Stability, Categories & Unseen UPCs

Testing generalization across time, store departments, and 606 cold-start products





Weekly & Category-Level Performance

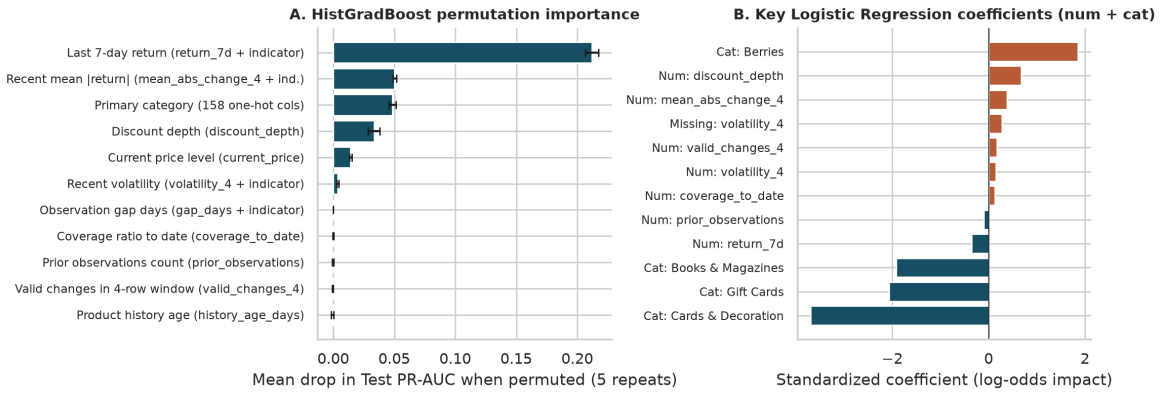
- Weekly Stability: HistGradBoost & Random Forest lead across all 3 test weeks (HistGB PR-AUC = 0.591 on Aug 23, 0.544 on Aug 30, 0.507 on Sep 6).
- Promotional Categories: Soda Pop and Juice-Non Refrigerated achieve Test F1 = 0.687 (PR-AUC = 0.865) and F1 = 0.654 (PR-AUC = 0.642).
- Static Categories: Cards & Decoration (0% change rate) produces 0 false positive alerts.

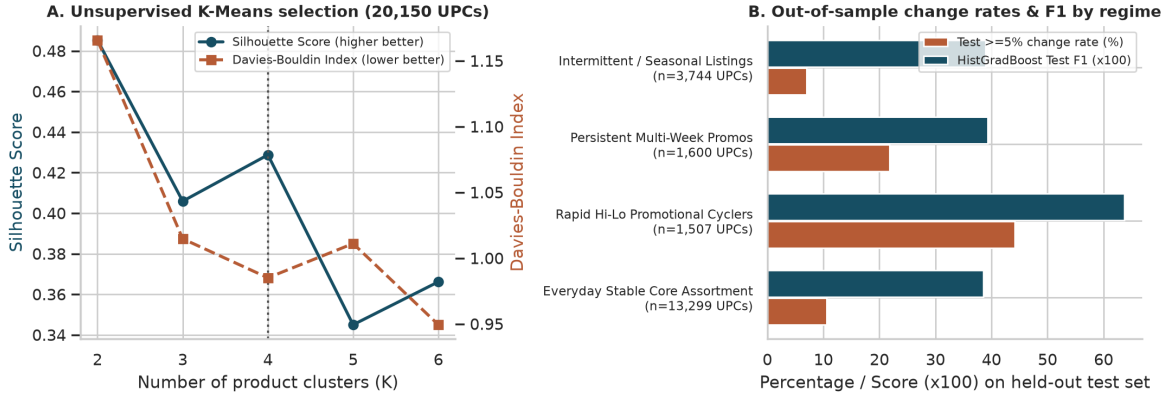
Seen vs. Cold-Start Unseen UPCs

- Seen UPCs (17,365 UPCs, 45,550 rows): ROC-AUC = 0.7894, PR-AUC = 0.5351, F1 = 0.4592, BalAcc = 0.7146.
- Unseen Cold-Start UPCs (606 new UPCs, 961 rows): ROC-AUC = 0.8548, BalAcc = 0.7223 using discount & category priors.

10. Feature Importance, Error Analysis & Unsupervised Regimes

What drives model predictions, where errors occur, and how K-Means product regimes align with predictability





Interpretability & Error Analysis

- Permutation Importance (PR-AUC Drop): Last 7d return (-0.212), Mean |return| (-0.050), Category (-0.049), Discount depth (-0.033).
- True Positives: Deep promotional discounts (e.g., 50% off FatBoy Ice Cream Sandwiches) rebounding after 1 week.
- False Positives / Negatives: Multi-week sales continuing into week 2 (FP) vs. unannounced new discounts on stable items (FN).

Unsupervised K-Means Regimes (20,150 Training UPCs)

- Cluster Selection: K=4 achieves local peak Silhouette = 0.429 & local minimum Davies-Bouldin = 0.985.
- Everyday Stable Core (13,299 UPCs): 5.7% discount freq, 97.7% coverage -> 10.6% test change rate.
- Rapid Hi-Lo Promos (1,507 UPCs): 23.7% mean |return| -> 44.1% test change rate (HistGB PR-AUC = 0.810, F1 = 0.637).

11. Model Recommendation, Conclusions & Future Directions

Deploying Histogram-Based Gradient Boosting for grocery price intelligence and next research steps

Final Recommendation: HistGradientBoostingClassifier

- Strongest Overall Trade-Off: Top-tier ranking (ROC-AUC 0.7935, PR-AUC 0.5364), fast 1.2s training, and non-linear interaction modeling.
- Flexible Deployment: Use tau* = 0.31 for F1-optimal alerts (F1 = 0.4785, Precision = 59.7%) or Balanced Weights for high Recall (59.0%, BalAcc = 71.5%).

Technical & Conceptual Lessons Learned

- Strict Temporal Hygiene: Enforcing target_date <= feature_date across train/test and CV folds prevents look-ahead leakage.
- Missingness as Signal: Binary missingness indicators distinguish 1-week vs. multi-week histories and resolve r=0.98 collinearity.
- Imbalance-Aware Evaluation: PR-AUC, F1, and Balanced Accuracy expose performance differences masked by 86% baseline accuracy.

Future Directions & Q&A Readiness

- 1. Cross-Retailer Entity Matching: Parse brand and package size strings to link identical products across King Kullen, Stop & Shop, and Gala Fresh.
- 2. Promotional Duration Features: Track consecutive discounted weeks ('promo_duration_weeks') to separate 1-week vs. multi-week sales and reduce False Positives.
- 3. Weekly Flyer NLP & Holiday Calendars: Integrate store circular text embeddings and major grocery holiday indicators.
- Q&A Topics Ready: Integer-cent target math, expanding-window CV vs. random CV, collinearity ablation, and cold-start UPC generalization.