GROCERY STORE RESEARCH: THE PRICE PULSE OF LONG ISLAND

Predicting Weekly Grocery Catalog Price Changes (>=5%) Across Longitudinal Retail Snapshots

Team Member / Author: Frank Valdez (Project Owner, Dataset Creator & Domain Reviewer) | Phase 3 Final Data Mining Project

1,031,264 — Raw Catalog Rows (82 Snapshots)

254,453 — Exact-7d Labeled King Kullen Pairs

0.794 / 0.543 — Peak Test ROC-AUC & PR-AUC

0.478 / 0.715 — Peak Test F1 & Balanced Acc.

1. Problem Statement & Research Objective

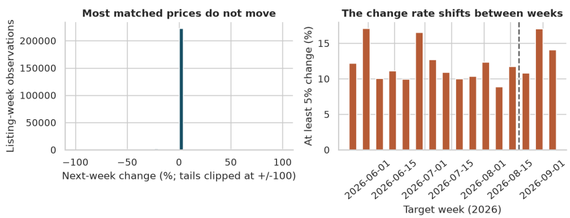
- Retail Grocery Problem: Supermarkets rely on High-Low ('Hi-Lo') promotional pricing—temporary weekly discounts followed by sharp rollbacks to regular list prices—creating 10%–50% weekly price swings for households.
- Research Question: What patterns in product attributes, observed discounts, and past price movements describe grocery catalogs, and can past-only features predict a >=5% price change exactly 7 days later?
- Exact Integer-Cent Target: y = 1 iff 100 * |future_cents(t+7) - current_cents(t)| >= 5 * current_cents(t). Unobserved future items remain unlabeled (never zero-filled).

2. Multi-Retailer Dataset & Cleaning Pipeline

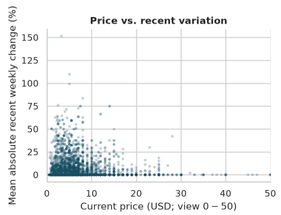
Retailer Archive	2026 Window	Raw	Clean	Median (IQR Fence)
King Kullen (Primary)	18 (May 16–Sep 6)	378,987	299,757	$4.99 ($15.04 fence)
Stop & Shop	3 (Jul 9–Jul 19)	127,002	127,002	$5.49 ($15.99 fence)
Gala Baldwin	16 (Jul 17–Aug 6)	149,505	149,505	$4.79 ($12.49 fence)
Gala Freeport	45 (Jul 23–Sep 6)	375,770	375,770	$4.59 ($11.74 fence)
Total Archive	82 Snapshots	1,031,264	952,034	161.58 MB Stored

- Consolidation: 79,230 repeated King Kullen category rows had 0 price conflicts and were merged into 299,757 unique observations by unioning category sets.
- Quality & Outliers: String UPCs preserve leading zeros; 51,861 evening UTC rows match NY dates; 915 items above the $15.04 fence are genuine bulk/catering goods (up to $249.00) and are retained.

3. Exploratory Data Analysis (EDA) Highlights







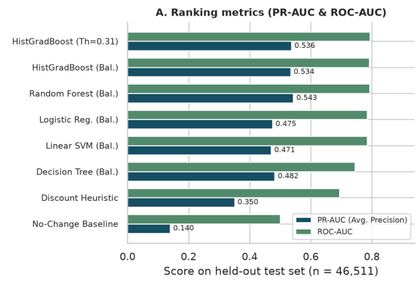
- Discount Effect: Items on promotional discount change price >=5% next week 42.04% of the time vs. 7.00% when undiscounted (+35.03 percentage points; 254,453 labeled pairs, 12.25% overall positive).

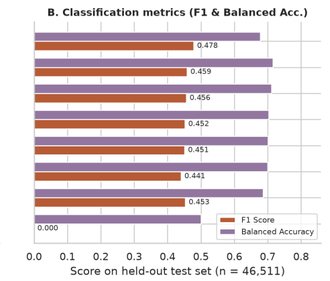
4. Methodology: Leak-Free Preprocessing & Temporal CV

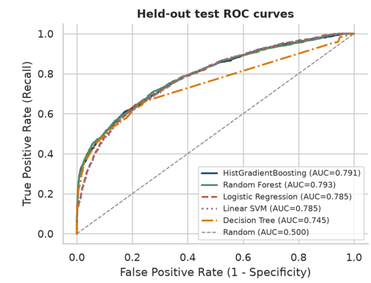
- Past-Only Feature Matrix (172 Columns): 10 numeric predictors (current_price, discount_depth, return_7d, mean_abs_change_4, volatility_4, valid_changes_4, gap_days, prior_observations, history_age_days, coverage_to_date) + 4 missingness indicators + 158 one-hot primary_category columns (pooled at min_frequency=100).
- Strict Chronological Allocation: Train on 13 target weeks (2026-05-24 to 2026-08-16; n=207,942, 11.87% pos); Hold out final 3 target weeks (2026-08-23 to 2026-09-06; n=46,511, 13.96% pos). Zero look-ahead leakage: max(train.target_date) <= min(test.date).
- 5-Fold Expanding-Window Temporal CV: Tuned hyperparameters across weeks 9–13 by refitting imputation, scaling, and encoding inside each fold.
- Collinearity Check (r = 0.98): Kept both mean_abs_change_4 (9.7% miss) and volatility_4 (19.2% miss) as their indicators distinguish 1 vs. >=2 prior changes (HistGB CV F1 = 0.4681 vs. 0.4633 if dropped).

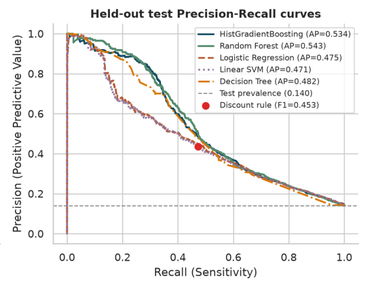
5. Key Results: Held-Out Test Set Evaluation (n = 46,511)

Model / Regime	Acc.	BalAcc	Prec.	Recall	F1	ROC	PR-AUC
No-Change Baseline (y=0)	0.8604	0.5000	0.0000	0.0000	0.0000	0.5000	0.1396
Observed Discount Rule	0.8408	0.6868	0.4352	0.4731	0.4534	0.6941	0.3503
Decision Tree (Balanced)	0.8029	0.6995	0.3648	0.5562	0.4406	0.7451	0.4821
Linear SVM (Balanced)	0.8164	0.7010	0.3871	0.5409	0.4513	0.7845	0.4706
Logistic Reg. (Balanced)	0.8139	0.7035	0.3837	0.5505	0.4522	0.7851	0.4746
Random Forest (Balanced)	0.8070	0.7120	0.3759	0.5802	0.4562	0.7929	0.5428
HistGradBoost (Balanced)	0.8057	0.7153	0.3753	0.5899	0.4588	0.7912	0.5340
HistGradBoost (Unweighted)	0.8891	0.6637	0.7066	0.3509	0.4690	0.7935	0.5364
HistGradBoost (Th=0.31)	0.8786	0.6777	0.5975	0.3990	0.4785	0.7935	0.5364

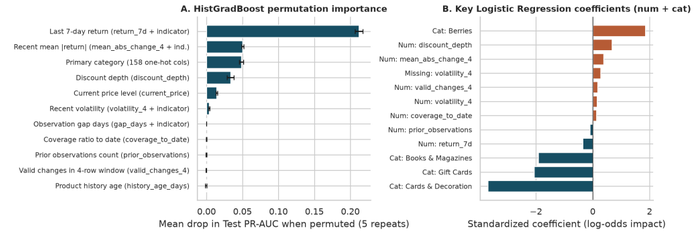


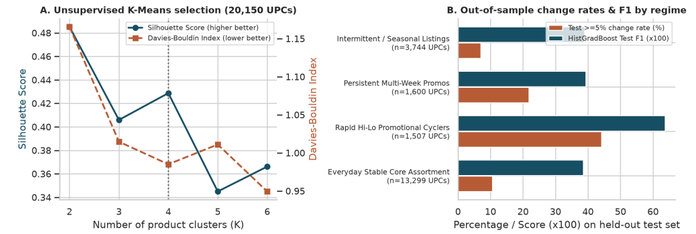






6. Feature Importance, Subgroups & Unsupervised Regimes





- Top Drivers (PR-AUC Drop): Last 7d return (-0.212), Mean |return| (-0.050), Primary category (-0.049), and Discount depth (-0.033).
- Unsupervised K-Means (20,150 UPCs): K=4 optimizes Silhouette (0.429) & Davies-Bouldin (0.985), separating Everyday Core Staples (13,299 UPCs, 10.6% test change) from Rapid Hi-Lo Promos (1,507 UPCs, 44.1% change, HistGB PR-AUC = 0.810, F1 = 0.637).

7. Discussion, Error Analysis & Lessons Learned

- Why Boosted Trees Excel: Grocery price shifts follow non-linear Hi-Lo rollback rules—an active discount (discount_depth > 0) paired with a recent price drop (return_7d < 0) signals an imminent return to regular price.
- Error Taxonomy: True Positives capture 1-week sales rebounding (e.g., 50% off FatBoy Ice Cream Sandwiches); False Positives occur when multi-week sales stay discounted into week 2 (e.g., Post Great Grains Cereal); False Negatives reflect unannounced new sales.
- Cold-Start Generalization: On 606 unseen test UPCs (961 rows), HistGB maintains ROC-AUC = 0.855 (BalAcc = 0.722) via discount depth and category priors.

8. Conclusion, Recommendation & Future Directions

- Recommended Model: HistGradientBoostingClassifier achieves top-tier ROC-AUC (0.7935) & PR-AUC (0.5364) in ~5s, offering F1 = 0.4785 (at CV threshold tau* = 0.31) or Balanced Accuracy = 0.7153 (with balanced weights).
- Future Work: (1) Cross-retailer product matching via brand/package-size parsing; (2) promotional duration ('weeks_on_sale') survival features to cut multi-week False Positives; (3) weekly circular NLP.