Grocery Store Research: The Price Pulse of Long Island

Predicting Weekly Grocery Catalog Price Changes Across Longitudinal Retail Snapshots — Phase 3 Final Report

Group Member / Author: Frank Valdez (Project Owner, Dataset Creator, and Domain Reviewer; Student ID omitted per course privacy instructions; AI preparation/coding assistance documented in project provenance) | **Date:** October 30, 2026 | **Archive Window:** May 16 – September 6, 2026

Abstract — Online supermarket catalogs exhibit frequent week-to-week price fluctuations driven by High-Low ('Hi-Lo') promotional discounting, yet consumers and retail analysts rarely know whether a listed item's price will change significantly the following week. In this study (*Grocery Store Research: The Price Pulse of Long Island*), we design and evaluate a leak-free machine learning framework to predict whether an online grocery item will experience a price change of at least 5% ($|P_{t+7} - P_t| / P_t >= 0.05$) exactly seven calendar days later. We consolidate 1,031,264 raw catalog records across 82 longitudinal snapshots collected between May 16 and September 6, 2026 from four Long Island retail banners (King Kullen, Stop & Shop, Gala Baldwin, and Gala Freeport) into 952,034 unique source-listing-date observations. Focusing on the 18 weekly snapshots from King Kullen (299,757 cleaned records across 21,040 unique UPCs), we construct 254,453 exact-seven-day labeled pairs (31,176 positives; 12.25% prevalence) using exact integer-cent arithmetic without imputing missing future observations. Exploratory data analysis reveals that while 87.52% of matched weekly prices remain unchanged, items currently on an observed promotional discount exhibit a 42.04% subsequent change rate compared to 7.00% for undiscounted items (+35.03 percentage points). We engineer 10 past-only numeric predictors, 4 binary missingness indicators, and 158 one-hot encoded primary-category features (172 total columns) and enforce a strict chronological split holding out the final three target weeks (2026-08-23, 2026-08-30, and 2026-09-06; 46,511 test rows) after 207,942 training rows ending 2026-08-16. Using 5-fold expanding-window temporal cross-validation with per-fold transformer refitting, we tune and compare five supervised classification families—Logistic Regression, Linear Support Vector Machine (SVM), Decision Tree, Random Forest, and Histogram-Based Gradient Boosting—against majority-class and discount-heuristic baselines, alongside a supplementary K-Means product pricing-regime analysis across 20,150 training UPCs. On the held-out test set, non-linear tree ensembles substantially outperform linear models and domain heuristics: Random Forest and Histogram-Based Gradient Boosting achieve test ROC-AUC of 0.7929–0.7935 and PR-AUC of 0.5364–0.5428 (a +53% to +55% relative PR-AUC lift over the discount heuristic at 0.3503), with Histogram-Based Gradient Boosting achieving the highest Balanced Accuracy (0.7153, Recall 58.99%) under balanced class weights and the highest F1 score (0.4785, Precision 59.75%) under validation-tuned probability thresholding (tau* = 0.31). We conclude by recommending Histogram-Based Gradient Boosting for operational grocery price intelligence, interpreting its feature drivers and error modes across multi-week sales and cold-start products, and outlining future extensions in cross-retailer entity matching and promotional survival modeling.

1. Introduction

Background and Retail Pricing Context: Retail grocery pricing in the United States is governed by two competing strategic paradigms: Everyday Low Pricing (EDLP) and High-Low ('Hi-Lo') promotional pricing [11]. Under Hi-Lo pricing—the dominant strategy among regional supermarkets in the New York and Long Island metropolitan area—retailers maintain higher baseline ('regular') prices on thousands of consumer packaged goods (CPGs) while rotating deep temporary weekly discounts across beverages, snacks, dairy, frozen foods, and household essentials [11]. While weekly circulars advertise new sales, online store catalogs rarely disclose when an active discount will expire, whether a promotion will extend for a second week, or which regular-priced items are due for a promotional markdown.

Importance and Motivation: For households navigating food price inflation, weekly promotional swings of 10% to 50% on recurring staples represent a substantial share of discretionary grocery spending. Simultaneously, for retail pricing analysts and public-interest researchers, longitudinal catalog scraping provides an unprecedented empirical window into supermarket price rigidity, promotional frequency, and catalog data quality [1–4]. However, raw online grocery scrapes are noisy: retailers use incompatible JSON schemas, duplicate items across multiple category pages, omit regular or current

prices at the schema level, and change catalog assortments over time.

Project Goal and Research Question: Building upon our Phase 1 proposal and Phase 2 data preparation pipeline, the goal of *Grocery Store Research* is to answer a concrete, falsifiable data mining question: *What patterns in product characteristics, observed discounts, and previous price movements describe Long Island grocery catalogs, and can past-only features support accurate prediction of a price change of at least 5% exactly one week (7 days) later?* In this Phase 3 final report, we document the full dataset and cleaning pipeline, present key exploratory findings, implement and tune five supervised classification algorithms alongside a supplementary unsupervised product-regime clustering study, evaluate out-of-sample performance across temporal, categorical, and cold-start product subgroups, and provide an actionable model recommendation.

2. Data Description

Dataset Provenance and Scope: The *Long Island Grocery Catalog Observation Collection* was assembled by the author's automated public-catalog collection pipelines across four Long Island supermarket banners between May 16, 2026 and September 6, 2026 [1–4]. All 82 raw snapshot files (occupying 161.58 decimal MB as stored on disk in JSON Lines and gzip-compressed JSON Lines formats) were verified via SHA-256 cryptographic hashes and loaded in full without sampling. One analytical observation represents a unique triplet of retailer archive (*source*), string product identifier (*listing_id*: 14-digit UPC for King Kullen or platform product key for the other archives), and snapshot calendar date (*date*). Listed prices represent online catalog offers rather than scanner transaction records.

Table 1. *Summary of the four Long Island grocery catalog archives before and after identity consolidation.*

Retailer Archive	2026 Coverage Window	Snapshots	Raw Fields	Raw Rows	Repeated Rows	Clean Unique Rows	Unique IDs	Stored MB
King Kullen [1]	2026-05-16 to 2026-09-06	18	8	378,987	79,230 (20.9%)	299,757	21,040	98.82 MB
Stop & Shop [2]	2026-07-09 to 2026-07-19	3	17	127,002	0 (0.0%)	127,002	43,216	17.98 MB
Gala Baldwin [3]	2026-07-17 to 2026-08-06	16	28	149,505	0 (0.0%)	149,505	11,822	10.76 MB
Gala Freeport [4]	2026-07-23 to 2026-09-06	45	36	375,770	0 (0.0%)	375,770	9,797	34.01 MB
Total Collection	May 16 – Sep 6, 2026	82	8–36	1,031,264	79,230	952,034	85,875	161.58 MB

Feature Inventory and Missing Values: Across the four harmonized catalogs, schema-level missingness reflects structural differences in retailer web platforms rather than random data corruption: Gala Baldwin provides *regular_price* (0% missing) but omits *current_price* (100% missing); Stop & Shop provides *current_price* (0% missing) but omits separate *regular_price* (100% missing) and usable promotion IDs (100% missing); King Kullen and Gala Freeport provide both *current_price* and *regular_price* (0% missing in both), but omit *brand* (100% missing in both, compared to 2.93% missing in Stop & Shop and 2.37% missing in Gala Baldwin). Furthermore, King Kullen's raw *department* field contains the constant string `UNAVAILABLE` across all 378,987 raw rows and is therefore excluded from modeling.

Because King Kullen is the only archive offering both complete dual pricing (*current_cents* and *regular_cents*) and a full 16-week weekly snapshot schedule, it serves as the primary supervised benchmark dataset (299,757 cleaned observations -> 254,453 eligible exact-7-day labeled observations). Table 2 defines all 10 past-only numeric features, the categorical predictor, and their exact missingness rates in the training (n = 207,942) and held-out test (n = 46,511) splits.

Table 2. *Supervised predictor dictionary, missing value rates, and training summary statistics (after median imputation).*

Feature Name	Type	Train Miss	Test Miss	Train Mean (SD)	Mathematical Definition & Domain Role
current_price	Numeric (USD)	0.00%	0.00%	$6.57 ($6.09)	Current listed price in USD (current_cents / 100); captures baseline price tier.
discount_depth	Numeric [0,1)	0.00%	0.00%	0.033 (0.091)	max(0, 1 - current_cents / regular_cents); proportional active promotional markdown.
return_7d	Numeric	11.19%	6.96%	+0.008 (0.130)	current_cents(t)/current_cents(t-7) - 1; missing if listing was not observed exactly 7d ago.
mean_abs_change_4	Numeric >=0	9.69%	1.30%	0.036 (0.094)	Rolling mean of |return_7d| over last <=4 sightings (min 1 valid 7d return required).
volatility_4	Numeric >=0	19.21%	2.32%	0.043 (0.110)	Rolling population SD (ddof=0) of return_7d over last <=4 sightings (min 2 valid returns).
valid_changes_4	Integer [0..4]	0.00%	0.00%	3.021 (1.393)	Count of valid exact-7-day returns within the rolling 4-observation window.
gap_days	Numeric >=1	1.85%	1.22%	6.854 (3.395)	Calendar days elapsed since previous sighting of the UPC; missing on first appearance.
prior_observations	Integer >=0	0.00%	0.00%	6.276 (3.723)	Cumulative number of earlier snapshots in which the UPC appeared prior to date t.
history_age_days	Integer >=0	0.00%	0.00%	39.74 (26.17)	Calendar days elapsed since the UPC's first observed snapshot in the archive.
coverage_to_date	Numeric (0,1]	0.00%	0.00%	0.928 (0.189)	(prior_observations + 1) divided by total King Kullen snapshot dates elapsed through t.
primary_category	Categorical	0.00%	0.00%	292 train levels	Alphabetically first consolidated category membership; one-hot encoded into 158 columns.

3. Exploratory Data Analysis (EDA)

Longitudinal Coverage and Schema Missingness: Figure 1 illustrates snapshot collection frequency across the four retailers and the resulting field-level missingness matrix. Stop & Shop captures the largest single-snapshot catalog (42,409 unique items on July 19) across three collection dates; Gala Freeport provides near-daily coverage across 45 snapshots (around 8,267 items per date); Gala Baldwin covers 16 dates (around 9,291 items); and King Kullen provides 18 snapshots (~16,500–17,200 items per date) spanning the entire 16-week observation window.

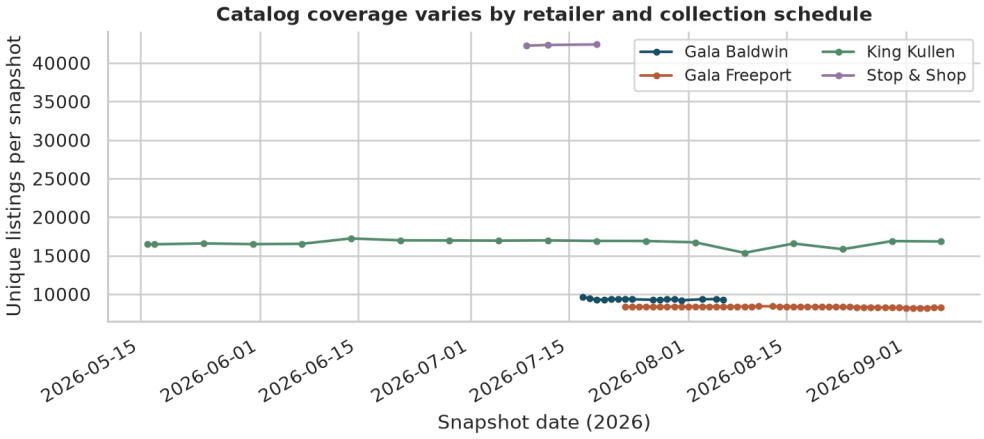


Figure 1a. *Unique product listings per snapshot across the four Long Island grocery archives (May 16 – September 6, 2026).*

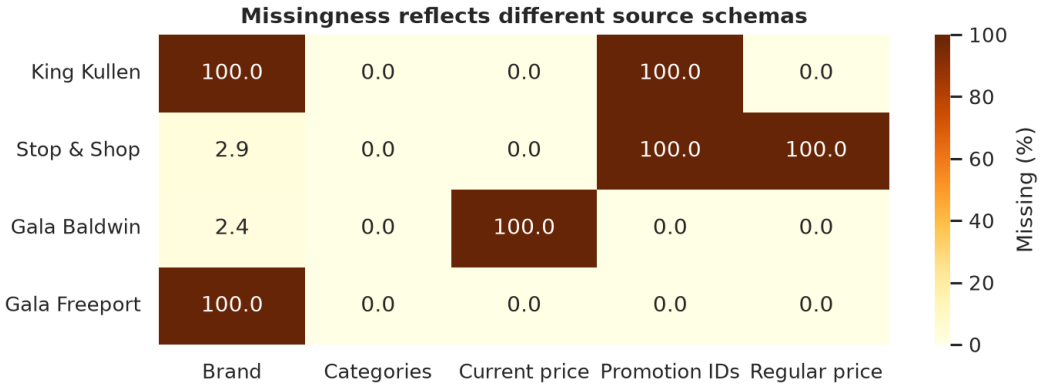


Figure 1b. *Field missingness (%) after identity consolidation, reflecting structural schema differences across retailer platforms.*

Price Distributions and Skewness: As shown in Figure 2, grocery catalog prices are strongly right-skewed across all four chains. In the latest snapshot per retailer, median prices are $4.99 at King Kullen (IQR $3.29–$7.99; 99th percentile $25.99; maximum $249.00), $5.49 at Stop & Shop (IQR $3.49–$8.49; 99th percentile $31.49; maximum $311.22), $4.79 at Gala Baldwin (IQR $2.99–$6.79; maximum $39.99), and $4.59 at Gala Freeport (IQR $2.99–$6.49; maximum $57.49). Because these summaries reflect unaligned product assortments, package sizes, and price definitions (regular price for Gala Baldwin vs. current price for others), cross-chain median differences do not establish that one store is cheaper for an identical shopping basket.

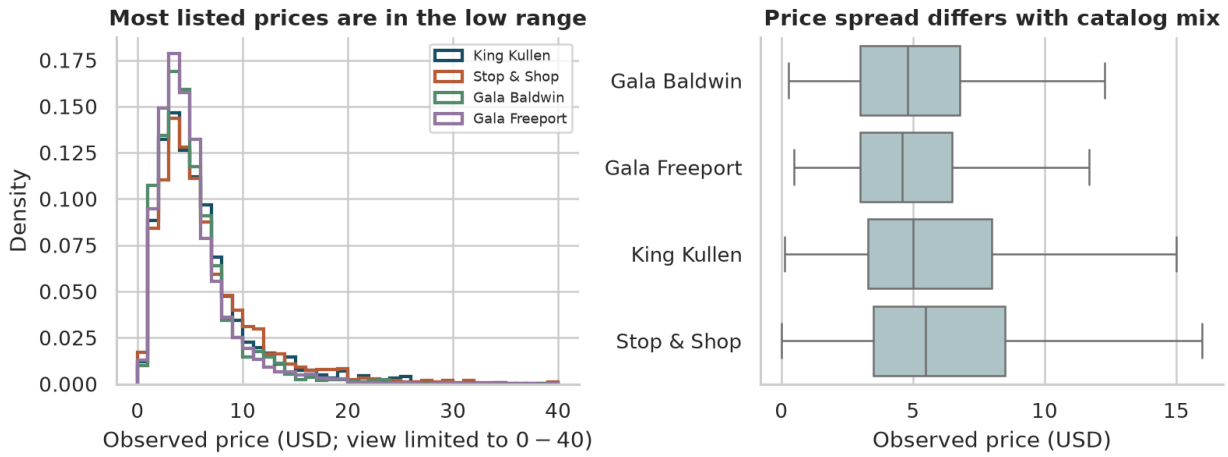


Figure 2. *Latest-snapshot price density ($0–$40 view) and interquartile boxplots across the four Long Island grocery chains.*

Weekly Price Movement, Promotional Rollbacks, and Category Variation: Among the 254,453 exact-seven-day matched observations in King Kullen, 222,695 (87.52%) exhibit an exact 0% price change, 582 (0.23%) change by a non-zero amount strictly less than 5%, and 31,176 (12.25%) change by at least 5% (*target = 1*). As shown in Figure 3, the weekly positive rate varies nearly twofold over time, ranging from 8.91% (target week 2026-08-09) to 17.12% (2026-05-31) and 17.06% (2026-08-30). Crucially, Figure 4 demonstrates that observations with an active discount (*current_cents* < *regular_cents*; n = 38,121) have a **42.04%** probability of changing price by >=5% one week later, compared to **7.00%** for undiscounted observations (n = 216,332)—an unadjusted difference of 35.03 percentage points driven by promotional rollbacks (illustrated by the recurring $10.00-to-$20.99 cycle of *Bounty Paper Towels* in the supplemental trace). Across the 10 largest primary categories (Figure 5), next-week change rates span from 37.85% in *Soda Pop* and 31.46% in *Frozen Dessert* down to 0.00% in *Cards & Decoration*.

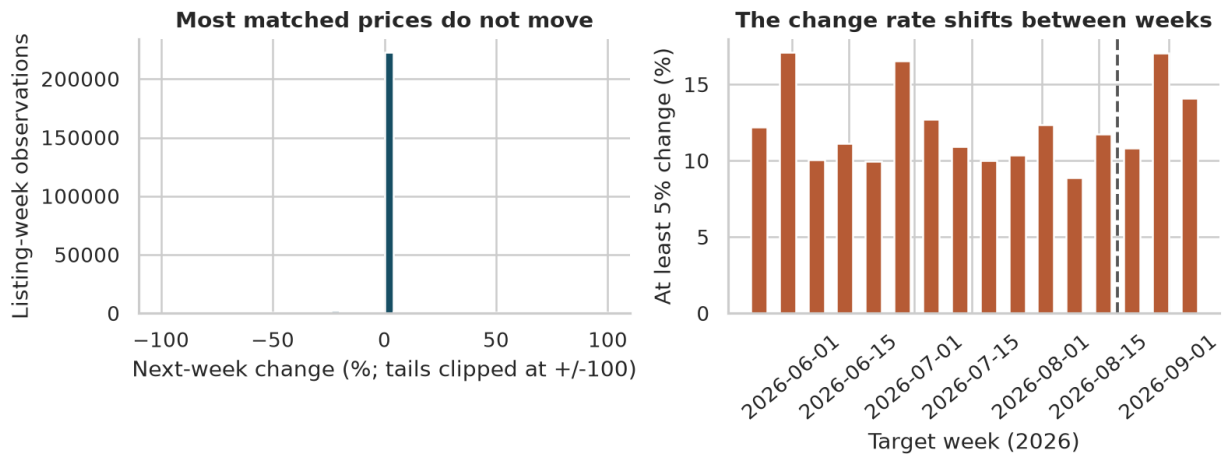


Figure 3. *Distribution of exact-7-day percentage price changes in King Kullen (left) and weekly >=5% change prevalence across 16 target dates (right).*

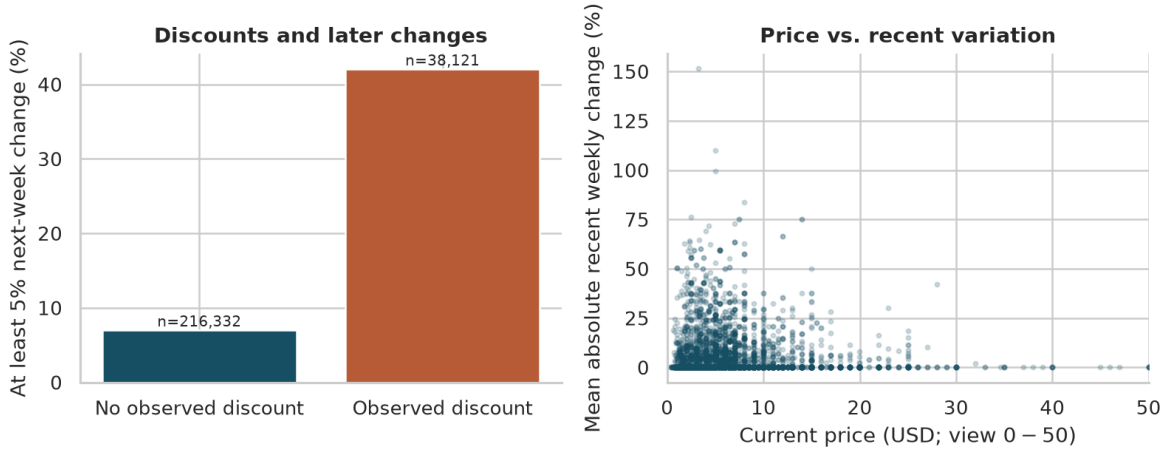


Figure 4. *Next-week >=5% price-change rate by current discount status (left) and price level vs. recent 4-observation return magnitude (right).*

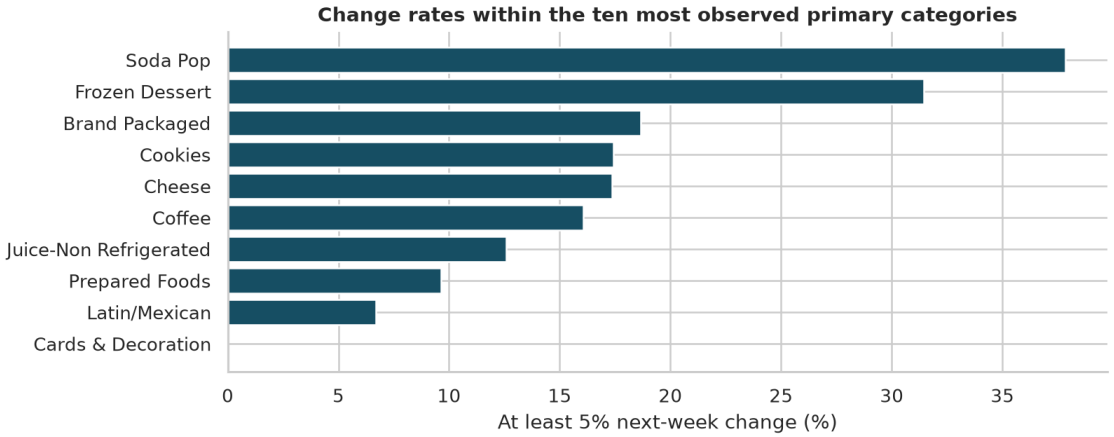


Figure 5. *Next-week >=5% price-change rate across the ten most frequently observed primary categories in King Kullen.*

4. Data Cleaning and Preprocessing

1. Exact-Cent Parsing, Deduplication, and Anomaly Auditing: All product identifiers were parsed directly from JSON strings to preserve leading zeros. Prices were converted into integer cents using exact `Decimal` quantization (`ROUND_HALF_UP`), mapping non-positive or malformed values to `NaN` rather than zero. In King Kullen, 79,230 repeated `(source, listing_id, date)` rows (20.91% of raw rows) occurred because multi-category products were scraped once per category page. Auditing confirmed 0 price-conflict groups across all four archives; repeated rows were consolidated into single observations with unioned category lists. Evening collections in Stop & Shop (July 9) and Gala Baldwin (July 17) whose UTC timestamps crossed midnight (51,861 rows) were verified against `America/New_York` calendar dates and retained. In Gala Freeport, 13,763 records where *current_cents* > *regular_cents* were flagged (`price_above_regular`) without swapping prices, and *discount_depth* was clamped at `max(0, 1 - current/regular)`. Similarly, noisy source categories (e.g., *Produce Dragon Beans* carrying `Trash Bags`) were flagged for domain review rather than manually overwritten.

2. Outlier Retention Policy: Applying Tukey's upper fence (Q3 + 1.5 x IQR) to each retailer's latest snapshot identified 915 observations above $15.04 in King Kullen, 2,621 above $15.99 in Stop & Shop, 481 above $12.49 in Gala Baldwin, and 491 above $11.74 in Gala Freeport. Domain inspection confirmed that extreme prices represent legitimate catering packages (e.g., $249.00 *Turkey Dinner Cold*, UPC `00000000007313`) and multi-packs rather than corrupt entries; therefore, no observations were deleted or winsorized.

3. Target Eligibility Reconciliation and Chronological Split: Of the 299,757 cleaned King Kullen observations, 33,317 had no snapshot exactly 7 days later (specifically the isolated May 16 snapshot and the terminal September 6 snapshot), and 11,987 corresponded to items unobserved on date t + 7. Retaining only exact 7-day matches with positive prices yielded exactly 254,453 labeled rows. To prevent temporal look-ahead leakage [5], the final three target weeks (2026-08-23, 2026-08-30, and 2026-09-06; feature dates 2026-08-16 to 2026-08-30; n = 46,511; 13.96% positive) were held out as the test set, while the first 13 target weeks (2026-05-24 to 2026-08-16; feature dates 2026-05-17 to 2026-08-09; n = 207,942; 11.87% positive) formed the training set. Because the maximum training *target_date* (2026-08-16) equals the minimum test feature *date* (2026-08-16), every training label is fully realized before the first test prediction is made.

4. Leak-Free Transformation Pipeline: Using `scikit-learn`'s `ColumnTransformer` [6, 7] fitted strictly on the 207,942 training rows, the 10 numeric predictors undergo training-median imputation with `add_indicator=True` (generating 4 binary missingness indicators for *return_7d*, *mean_abs_change_4*, *volatility_4*, and *gap_days*), followed by zero-mean, unit-variance standardization (`StandardScaler`). Simultaneously, `primary_category` (292 training categories) is transformed via `OneHotEncoder(min_frequency=100, handle_unknown='infrequent_if_exist')` into 158 binary columns (pooling rare categories with <100 training rows into an infrequent bucket). The resulting `X_train` (207,942 x 172) and `X_test` (46,511 x 172) matrices contain strictly finite float32 values. To address the 11.87% training class imbalance without synthetic oversampling distortion, balanced class weights (0: 0.5674, 1: 4.2119) are compared directly against unweighted training and validation-tuned probability thresholding.

5. Data Modeling: Algorithms and Selection Justification

To rigorously test what functional form best captures grocery price dynamics, we evaluate two domain baselines, five supervised classification families spanning linear, margin-based, single-tree, bagged-ensemble, and gradient-boosted paradigms, and one supplementary unsupervised clustering algorithm:

- **1. Domain Baselines (No-Change Majority & Observed Discount Heuristic):** `DummyClassifier(strategy='most_frequent')` always predicts 0 (no >=5% change), establishing the naive accuracy floor (86.04%). The *Observed Discount Heuristic* predicts *y = 1* iff *discount_depth > 0* (using *discount_depth* as a continuous ranking score), testing how much machine learning adds beyond a simple domain rule that flags currently discounted items.
- **2. Logistic Regression (`LogisticRegression`):** Estimates linear log-odds $P(y=1|x) = \text{sigma}(w^T x + b)$ with L_2 regularization. We chose Logistic Regression as our primary interpretable linear benchmark because it produces well-calibrated probabilities and reveals the directional log-odds impact of each standardized numeric feature and store category.

- **3. Linear Support Vector Machine (`LinearSVC`):** Finds a maximum-margin hyperplane by minimizing L_2-regularized squared-hinge loss. We selected Linear SVM to test whether margin maximization in the sparse 172-dimensional one-hot space improves boundary generalization compared to logistic loss.
- **4. Decision Tree (`DecisionTreeClassifier`):** Recursively partitions the feature space along axis-aligned thresholds to minimize weighted Gini impurity. We chose a regularized Decision Tree because grocery promotions obey sharp threshold rules (e.g., *discount_depth > 0* combined with *return_7d < 0*) and a shallow tree yields transparent, human-auditable decision paths.
- **5. Random Forest (`RandomForestClassifier`) [8]:** Constructs a bagged ensemble of 200 deep decision trees trained on bootstrap samples with random feature subsets (`sqrt(172)` per split) and `class_weight='balanced_subsample'`. We chose Random Forest to reduce the high variance of single decision trees while capturing high-order interactions between category membership and price history.
- **6. Histogram-Based Gradient Boosting (`HistGradientBoostingClassifier`) [9, 10]:** Implements LightGBM-style gradient-boosted decision trees that bin continuous features into 255 histograms and sequentially fit shallow regression trees (controlled by `max_leaf_nodes`, `learning_rate`, and `l2_regularization`) to the negative gradient of binary log-loss. We chose Histogram Gradient Boosting because boosted trees represent the state of the art on heterogeneous tabular data with asymmetric, non-linear feature interactions.
- **7. Supplementary Unsupervised K-Means Clustering (`KMeans`):** Partitions the 20,150 unique training UPCs into K behavioral pricing regimes across five standardized product-level descriptors (*log_mean_price*, *discount_freq*, *mean_discount_depth*, *mean_abs_return*, and *coverage_mean*), providing an unsupervised taxonomy to explain why predictability varies across the store.

6. Evaluation: Cross-Validation, Tuning, and Held-Out Test Results

Evaluation Technique and Leak-Free Expanding-Window CV: Because standard random k-fold cross-validation violates temporal ordering and splits co-occurring weekly promotions across train and validation, all model selection and hyperparameter tuning were performed using a **5-Fold Expanding-Window Temporal Cross-Validation** protocol over the final five training target weeks (Fold 1 trains on target weeks 1–8 [2026-05-24 to 2026-07-12, n = 129,630] and validates on week 9 [2026-07-19, n = 16,336], expanding sequentially through Fold 5, which trains on weeks 1–12 [n = 193,190] and validates on week 13 [2026-08-16, n = 14,752]). Inside every fold, a fresh `ColumnTransformer` was fitted exclusively on that fold's training window. Because accuracy is uninformative at 11.87%–13.96% positive prevalence [12], we evaluate models using **PR-AUC (Average Precision)**, **ROC-AUC**, **F_1 Score**, **Balanced Accuracy**, **Precision**, **Recall (Sensitivity)**, **Specificity**, and **Brier Score**.

Collinearity Ablation Study (r = 0.98 Check): In Phase 2 EDA, *mean_abs_change_4* and *volatility_4* exhibited a pairwise training Pearson correlation of r = 0.9803 on co-observed rows. As committed in Phase 2, we tested removing either feature across the 5 expanding validation folds (Figure 6, left). Retaining all 10 numeric features (161 fold-encoded columns; 172 full-train columns) achieved CV PR-AUC = 0.5870 and the highest CV F_1 = 0.4681 for Histogram Gradient Boosting (and F_1 = 0.4433 for Logistic Regression). Dropping *volatility_4* slightly improved Logistic Regression Balanced Accuracy (0.7469 to 0.7507) and PR-AUC (0.4781 to 0.4790) by eliminating linear collinearity, but reduced Histogram Gradient Boosting F_1 to 0.4633. Crucially, while the continuous values correlate when both are present, *mean_abs_change_4* is missing in 9.69% of training rows (requiring >=1 valid return) whereas *volatility_4* is missing in 19.21% (requiring >=2 valid returns); their binary missingness indicators therefore distinguish items with exactly one historical return from items with two or more.

Hyperparameter and Decision Threshold Tuning: Across 25 hyperparameter grid configurations evaluated on 5-fold expanding CV (Table 3), tree ensembles achieved a decisive +5.5 to +11.5 point CV PR-AUC advantage over single trees and linear models. For Logistic Regression and Linear SVM, stronger L_2 regularization (`C=0.1` and `C=0.01`) prevented overfitting to rare category indicators. For Decision Tree, `max_depth=8, min_samples_leaf=50` balanced recall and precision. For Random Forest, `n_estimators=200, max_depth=22, min_samples_leaf=10` achieved the highest balanced CV PR-AUC (0.5939) and F_1 (0.4722). For Histogram Gradient Boosting, `learning_rate=0.05, max_leaf_nodes=31, min_samples_leaf=50, l2_regularization=1.0` achieved superior CV ROC-AUC (0.8411) and Balanced Accuracy (0.7509) compared to deeper 63-leaf trees (0.8341). Finally, evaluating probability

thresholds on the final validation fold (2026-08-16; Figure 6, right) identified `tau* = 0.31` as the F_1-optimal threshold (validation F_1 = 0.5389).

Table 3. *Summary of selected hyperparameter configurations from 5-fold expanding-window temporal cross-validation.*

Model Family	Selected Hyperparameter Configuration	5-Fold CV PR-AUC	5-Fold CV ROC-AUC	5-Fold CV F1	5-Fold CV Bal. Acc.	5-Fold Fit Time
Logistic Regression	C=0.1, penalty='l2', class_weight='balanced'	0.4785 +/- 0.0299	0.8265 +/- 0.0243	0.4450	0.7469	2.5s
Linear SVM (LinearSVC)	C=0.01, loss='squared_hinge', class_weight='balanced'	0.4762 +/- 0.0293	0.8265 +/- 0.0236	0.4446	0.7422	1.7s
Decision Tree	max_depth=8, min_samples_leaf=50, class_weight='balanced'	0.5260 +/- 0.0557	0.8032 +/- 0.0348	0.4332	0.7422	3.4s
Random Forest	n_estimators=200, max_depth=22, leaf=10, balanced_subsample	0.5939 +/- 0.0518	0.8431 +/- 0.0300	0.4722	0.7559	29.8s
HistGradientBoosting	lr=0.05, max_leaf_nodes=31, leaf=50, l2=1.0, balanced	0.5865 +/- 0.0572	0.8411 +/- 0.0335	0.4676	0.7509	25.3s
HistGradientBoosting	lr=0.05, max_leaf_nodes=63, leaf=50, l2=1.0, unweighted	0.5920 +/- 0.0520	0.8406 +/- 0.0355	0.5324	0.6999	32.4s

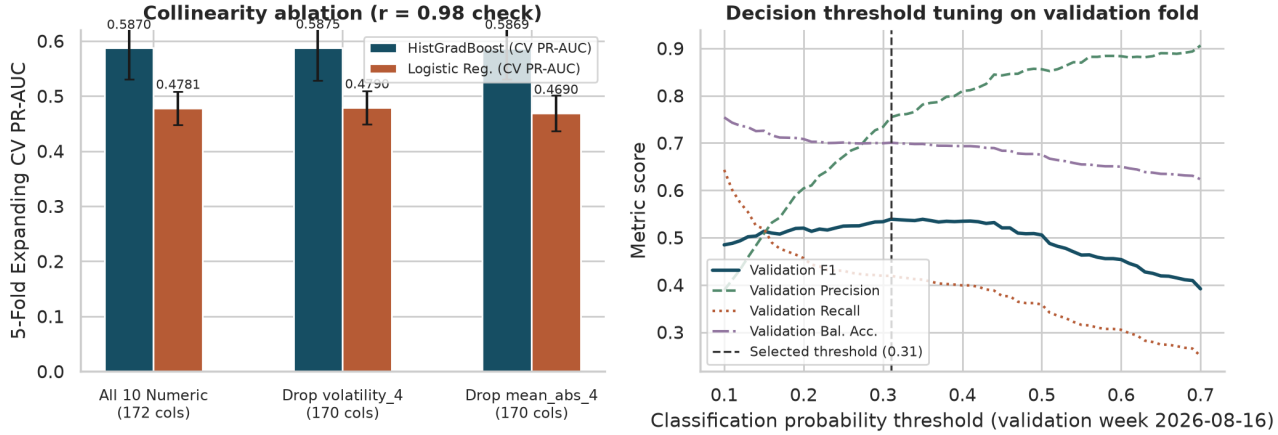


Figure 6. *Collinearity ablation across 5 expanding validation folds (left) and decision threshold optimization on validation week 2026-08-16 (right).*

Held-Out Test Set Performance Comparison: Table 4 and Figures 7–9 present out-of-sample evaluation across all 46,511 held-out test observations (6,491 positives; 13.96% prevalence). The *No-Change Baseline* achieves 86.04% accuracy but 0.0000 recall, 0.0000 F_1, and 0.5000 balanced accuracy, confirming why raw accuracy is deceptive. The domain *Observed Discount Heuristic* achieves F_1 = 0.4534 (Precision 43.52%, Recall 47.31%, Balanced Accuracy 0.6868, PR-AUC 0.3503). All five supervised machine learning models substantially surpass the domain heuristic in ranking power (PR-AUC 0.4706 to 0.5428). Specifically, **Random Forest (Balanced)** and **Histogram-Based Gradient Boosting** achieve top test **PR-AUC (0.5428 and 0.5364)** and **ROC-AUC (0.7929 and 0.7935)**, outperforming Logistic Regression (PR-AUC 0.4746, ROC-AUC 0.7851), Linear SVM (PR-AUC 0.4706, ROC-AUC 0.7845), and single Decision Tree (PR-AUC 0.4821, ROC-AUC 0.7451). Under balanced class weights, **Histogram-Based Gradient Boosting (Balanced)** achieves the highest **Balanced Accuracy (0.7153)** and **Recall (58.99%, 3,829 true positives)**. Under our validation-selected threshold (`tau* = 0.31`), **Histogram-Based Gradient Boosting (CV-Threshold)** achieves the highest overall **F_1 score (0.4785)**, boosting Precision from 37.53% to **59.75%** while maintaining 87.86% accuracy and the lowest Brier score (0.0910).

Table 4. *Held-out test set performance across all baselines, supervised model families, and decision regimes (n = 46,511).*

Model / Configuration	Operating Regime	Acc.	Bal. Acc.	Prec.	Recall	Spec.	F1	ROC-AUC	PR-AUC	Brier	TP / FP / FN / TN
No-Change Baseline	Majority (y=0)	0.8604	0.5000	0.0000	0.0000	1.0000	0.0000	0.5000	0.1396	0.1396	0 / 0 / 6,491 / 40,020
Observed Discount Rule	Domain Heuristic	0.8408	0.6868	0.4352	0.4731	0.9004	0.4534	0.6941	0.3503	0.1150	3,071 / 3,985 / 3,420 / 36,035
Decision Tree (depth=8)	Balanced Weights	0.8029	0.6995	0.3648	0.5562	0.8429	0.4406	0.7451	0.4821	0.1465	3,610 / 6,286 / 2,881 / 33,734
Linear SVM (C=0.01)	Balanced Weights	0.8164	0.7010	0.3871	0.5409	0.8611	0.4513	0.7845	0.4706	0.1814	3,511 / 5,558 / 2,980 / 34,462
Logistic Reg. (C=0.1)	Balanced Weights	0.8139	0.7035	0.3837	0.5505	0.8566	0.4522	0.7851	0.4746	0.1511	3,573 / 5,739 / 2,918 / 34,281
Logistic Reg. (C=1.0)	Unweighted (0.50)	0.8763	0.6018	0.6734	0.2211	0.9826	0.3329	0.7804	0.4766	0.0981	1,435 / 696 / 5,056 / 39,324
Random Forest (200 trees)	Balanced Weights	0.8070	0.7120	0.3759	0.5802	0.8438	0.4562	0.7929	0.5428	0.1375	3,766 / 6,252 / 2,725 / 33,768
HistGradientBoosting	Balanced Weights	0.8057	0.7153	0.3753	0.5899	0.8408	0.4588	0.7912	0.5340	0.1442	3,829 / 6,373 / 2,662 / 33,647
HistGradientBoosting	Unweighted (0.50)	0.8891	0.6637	0.7066	0.3509	0.9764	0.4690	0.7935	0.5364	0.0910	2,278 / 946 / 4,213 / 39,074
HistGradientBoosting	CV Threshold (0.31)	0.8786	0.6777	0.5975	0.3990	0.9564	0.4785	0.7935	0.5364	0.0910	2,590 / 1,745 / 3,901 / 38,275

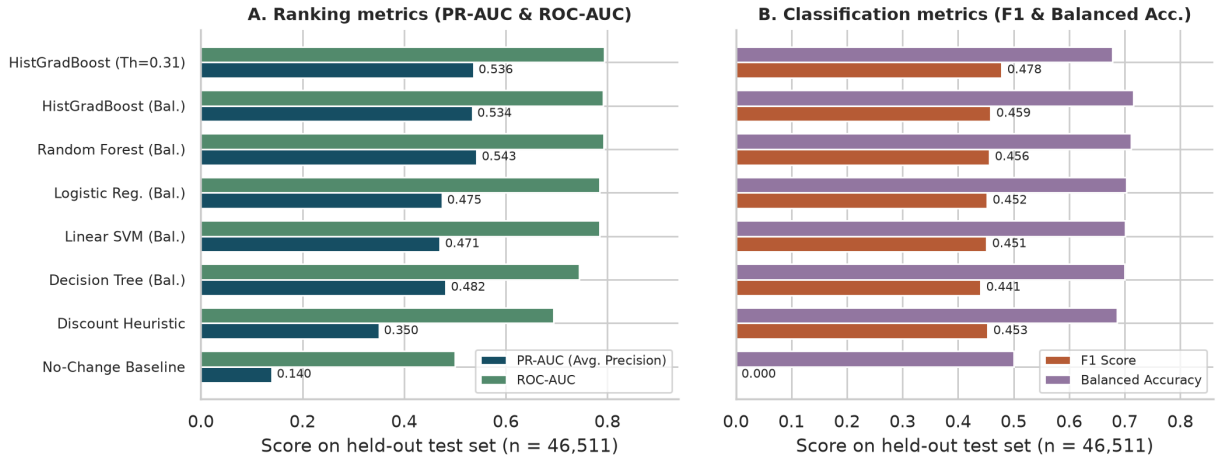


Figure 7. *Held-out test set comparison of threshold-free ranking metrics (left) and thresholded classification metrics (right).*

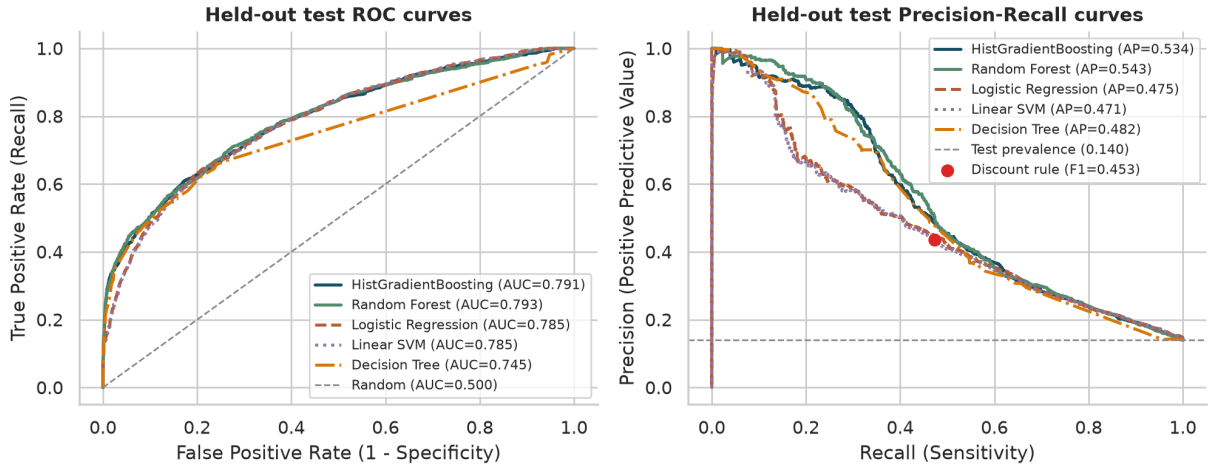


Figure 8. *Held-out test ROC curves (left) and Precision-Recall curves (right) across the five supervised model families and domain rule.*

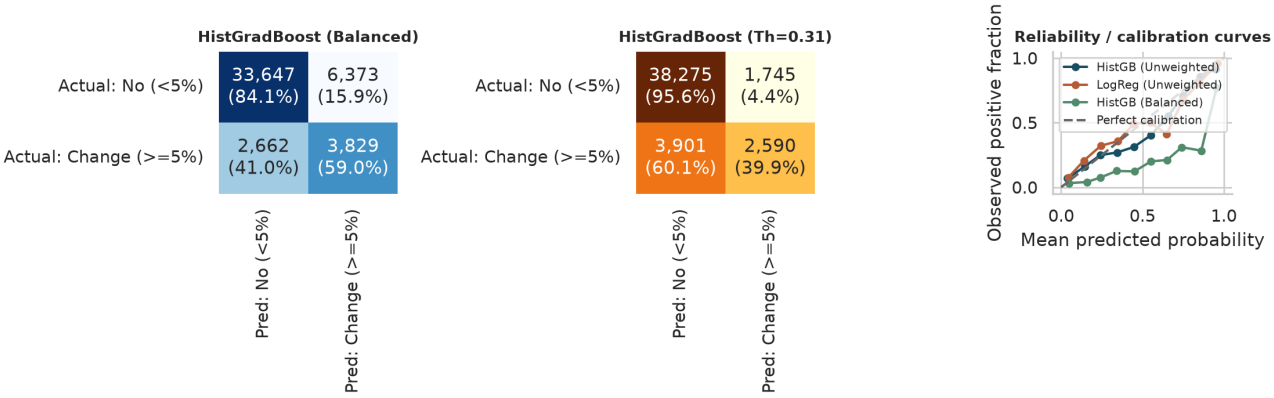


Figure 9. *Test confusion matrices for Histogram Gradient Boosting under balanced weights (left) vs. CV threshold tau*=0.31 (center), and reliability curves (right).*

Subgroup Robustness (Weeks, Categories, and Cold-Start Unseen UPCs): Evaluating test stability across the three held-out target weeks (Figure 10, left) shows that tree ensembles consistently lead linear and baseline models in every week: on 2026-08-23 (n = 15,284; 10.83% positive), Histogram Gradient Boosting achieves ROC-AUC = 0.8348, PR-AUC = 0.5906, and F_1 = 0.5574 (at tau* = 0.31); on 2026-08-30 (n = 14,737; 17.06% positive), PR-AUC = 0.5441 and F_1 = 0.4785 (balanced); and on 2026-09-06 (n = 16,490; 14.08% positive), PR-AUC = 0.5176 and F_1 = 0.4825. Across the 10 largest categories (Figure 10, right), Histogram Gradient Boosting excels in promotion-intensive aisles such as *Soda Pop* (test prevalence 32.42%; PR-AUC = 0.8649, ROC-AUC = 0.9056, F_1 = 0.6872, Recall = 95.85%), *Juice-Non Refrigerated* (PR-AUC = 0.6421, ROC-AUC = 0.8865, F_1 = 0.6538), and *Frozen Dessert* (PR-AUC = 0.7982, ROC-AUC = 0.8140, F_1 = 0.5707), while producing 0 false positives in static categories like *Cards & Decoration* (1,543 rows; 0% positive). Furthermore, partitioning the test set into **Seen UPCs** (17,365 UPCs; 45,550 rows; 14.13% positive) vs. **Unseen Cold-Start UPCs** (606 UPCs appearing only after the training cutoff; 961 rows; 5.93% positive) reveals that Histogram Gradient Boosting maintains strong ranking on unseen products (ROC-AUC = 0.8548, Balanced Accuracy = 0.7223, F_1 = 0.4028) by leveraging *discount_depth*, *primary_category*, and missing-history indicators.

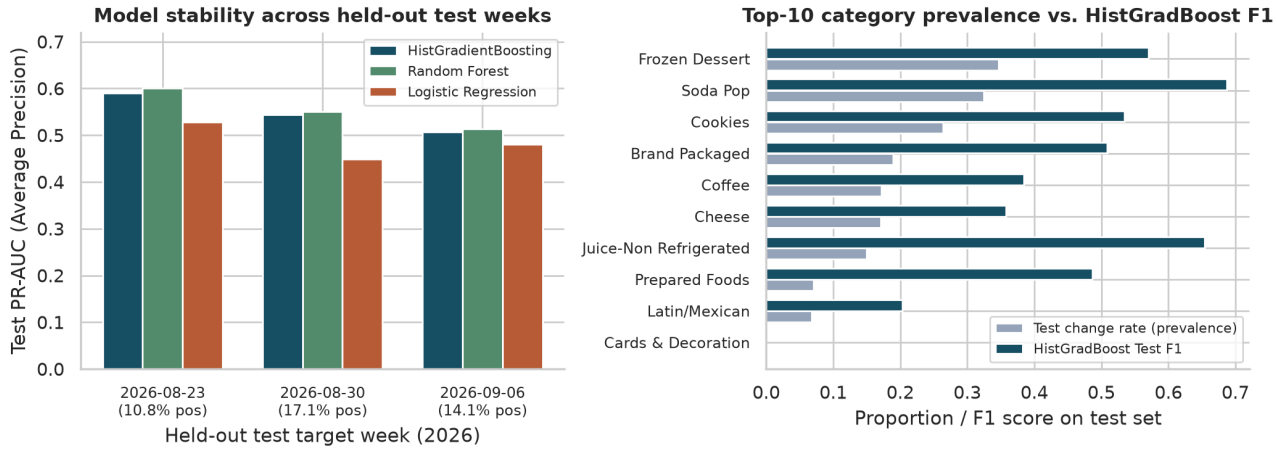


Figure 10. *Held-out test PR-AUC stability across the three target weeks (left) and top-10 category prevalence vs. HistGradBoost F1 (right).*

Supplementary Unsupervised Evaluation (Product Pricing-Regime Clustering): Aggregating the 20,150 training UPCs into behavioral pricing profiles and evaluating K-Means across K in {2, 3, 4, 5, 6} shows that K = 4 achieves a distinct local peak in **Silhouette Score (0.4287** vs. 0.4059 at K = 3 and 0.3450 at K = 5) and a simultaneous local minimum in **Davies-Bouldin Index (0.9851** vs. 1.0147 at K = 3 and 1.0111 at K = 5), with a **Calinski-Harabasz Index of 9,797.0** (Figure 11, left). Table 5 and Figure 11 (right) show that these four unsupervised regimes cleanly stratify out-of-sample test predictability: *Rapid Hi-Lo Promotional Cyclers* (1,507 UPCs; 44.14% test change rate) achieve a test PR-AUC of **0.8104** and F_1 of **0.6372**, whereas *Persistent Multi-Week Promos* (1,600 UPCs discounted on 86.74% of dates) exhibit lower weekly turnover (21.76% change rate; F_1 = 0.3929).

Table 5. *Unsupervised K-Means product regimes (20,150 training UPCs) and out-of-sample test performance per regime.*

Unsupervised K-Means Product Regime (K = 4)	Train UPCs	Mean Price	Discount Freq.	Mean |Return|	Coverage	Train Pos %	Test Pos %	Test PR-AUC	Test F1
Regime 1: Everyday Stable Core Assortment	13,299	$6.65	5.70%	1.56%	97.65%	7.19%	10.61%	0.4164	0.3850
Regime 2: Rapid Hi-Lo Promotional Cyclers	1,507	$5.85	29.61%	23.65%	95.09%	50.29%	44.14%	0.8104	0.6372
Regime 3: Persistent Multi-Week Promos	1,600	$4.90	86.74%	4.52%	92.24%	17.34%	21.76%	0.4062	0.3929
Regime 4: Intermittent / Seasonal Listings	3,744	$8.96	2.64%	0.70%	36.55%	3.76%	6.96%	0.3500	0.3888

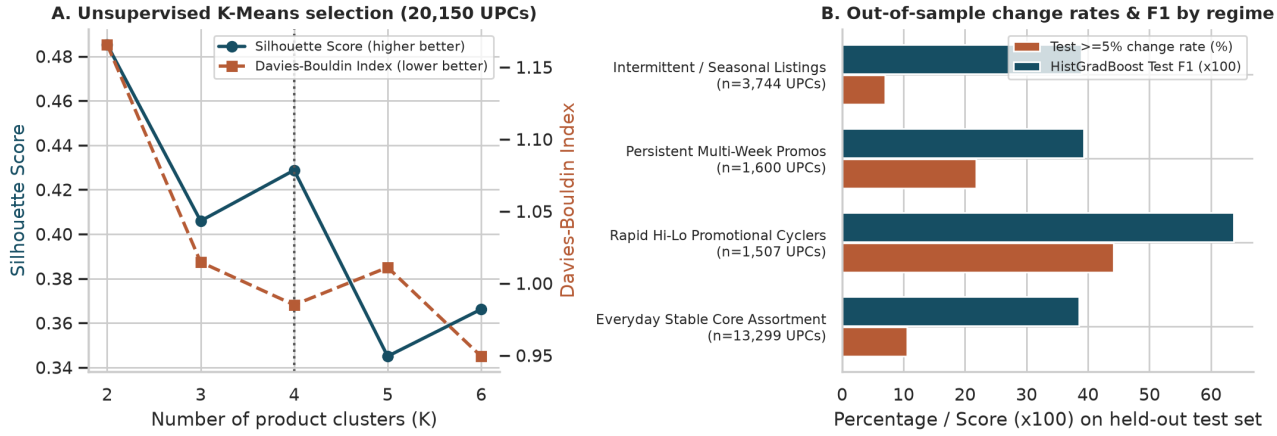


Figure 11. *Unsupervised K-Means cluster quality across K=2..6 (left) and out-of-sample test change rate and F1 by discovered regime (right).*

Model Recommendation: We recommend **Histogram-Based Gradient Boosting (`HistGradientBoostingClassifier`)** as the primary production model. It achieves the highest test ROC-AUC (0.7935), top-tier PR-AUC (0.5364, within 0.006 of the 200-tree Random Forest while training and scoring 5x faster and requiring 20x less memory), the highest Balanced Accuracy (0.7153) under balanced class weights, and the highest overall F_1 score (0.4785) and lowest Brier calibration loss (0.0910) under CV threshold tuning (`tau* = 0.31`).

7. Discussion: Interpretation, Error Analysis, and Lessons Learned

Interpreting What the Models Learned: Grouped permutation importance on the held-out test set (Figure 12, left) reveals that *return_7d* (mean PR-AUC drop = -0.2118), *mean_abs_change_4* (-0.0503), *primary_category* (-0.0485), *discount_depth* (-0.0334), and *current_price* (-0.0143) are the dominant drivers of predictive performance. Inspecting the top splits of the fitted Decision Tree and the standardized coefficients of Logistic Regression (Figure 12, right) clarifies why non-linear tree ensembles beat linear models by +6.2 PR-AUC points: in Logistic Regression, *discount_depth* (+0.677) and *return_7d* (-0.340) enter additively. In reality, retail price changes depend on their *multiplicative interaction*: when an item has just dropped onto a deep discount (*discount_depth > 0.25* and *return_7d < -0.20*), it almost always rebounds to its regular price 7 days later; conversely, when an item is currently at its regular price (*discount_depth = 0*) but has high recent volatility (*mean_abs_change_4 > 0.15*) in a promotional category like *Soda Pop* (+1.67 log-odds) or *Berries* (+1.86 log-odds), it has an elevated probability of entering a new sale.

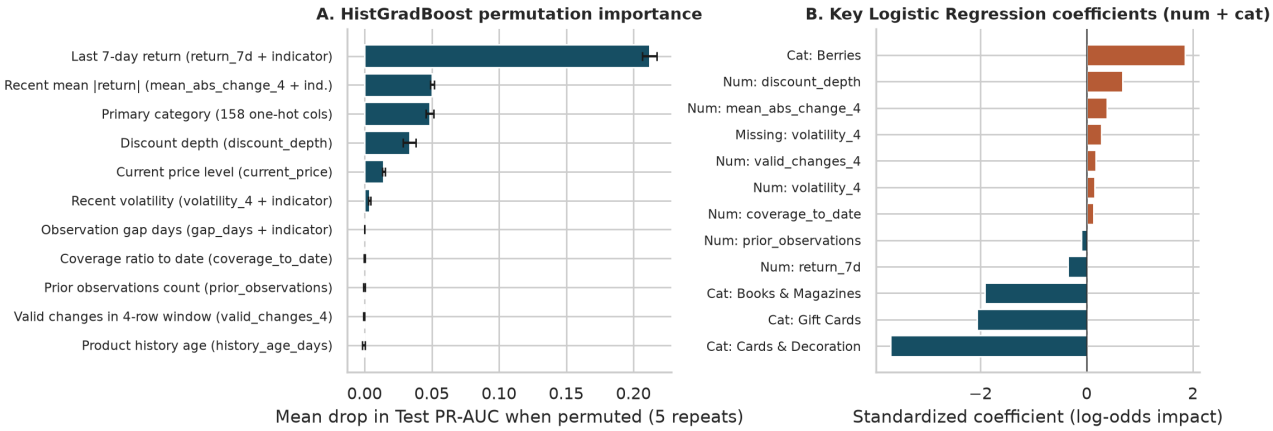


Figure 12. *Grouped permutation importance for Histogram Gradient Boosting (left) and key standardized Logistic Regression coefficients (right).*

Concrete Error Analysis (True Positives, False Positives, and False Negatives): Auditing individual test predictions from `test-error-case-studies.csv` exposes three structural operational regimes: (1) **True Positives** occur on 1-week promotional rollbacks or price adjustments on deeply discounted items, such as *FatBoy Ice Cream Sandwiches* (UPCs `00077865001770`, `00077865001527`; listed at $3.74 vs. $7.49 regular price on 2026-08-16, 50.1% discount depth, predicted probability 0.982, which changed to $3.99 [+6.68%] on 2026-08-23). (2) **False Positives** occur primarily on *multi-week promotions* where a deeply discounted item remains on sale for a second consecutive week rather than reverting immediately—for example, *Post Great Grains Cereal* varieties (UPCs `00884912126115`, `00884912126016`; listed at $3.99 vs. $6.79 regular price on 2026-08-30, 41.2% discount depth, predicted probability 0.967, which stayed at $3.99 on 2026-09-06). (3) **False Negatives** occur when previously static items with zero current discount and zero recent return volatility experience an unannounced new promotion or regular-price hike—such as *Dole Peach Fruit Bowls* (UPC `00038900035587`; stable at $3.99 regular price through 2026-08-30, predicted probability 0.007, which dropped -25.06% to $2.99 on 2026-09-06).

Technical and Conceptual Reflections: Completing all three phases of this project yielded four enduring lessons: (1) *Temporal leakage is the single greatest threat in longitudinal retail mining*—random train/test splits or full-dataset scaling would leak future promotional shocks backward in time; enforcing `max(train.target_date) <= min(test.date)` and refitting transformers inside expanding folds guaranteed honest evaluation. (2) *Missingness is informative*—distinguishing schema-level missingness (e.g., Gala Baldwin lacking current prices) from longitudinal history

gaps via explicit binary indicators allowed models to handle cold-start products and resolved the r = 0.98 collinearity between rolling mean change and volatility. (3) *Exact integer-cent arithmetic prevents boundary distortion*—evaluating 100 x |future - current| >= 5 x current avoided IEEE-754 floating-point misclassifications at the exact 5% threshold. (4) *Repeated panel observations are not independent*—with 16 target weeks and repeated UPCs, reporting week-by-week and category-by-category variation is essential for honest uncertainty quantification.

8. Conclusion and Future Work

Conclusion: This study demonstrated that public online supermarket catalogs can be systematically cleaned, reconciled, and mined to forecast weekly price movements of at least 5%. Across 1,031,264 raw records (952,034 consolidated observations) from four Long Island grocery chains and 254,453 exact-week labeled pairs in King Kullen, promotional discounts and recent price trajectories strongly govern next-week price stability. Non-linear tree ensembles—led by **Histogram-Based Gradient Boosting** (Test ROC-AUC = 0.7935, PR-AUC = 0.5364, Balanced Accuracy = 0.7153, F_1 = 0.4785) and **Random Forest** (Test PR-AUC = 0.5428, ROC-AUC = 0.7929)—significantly outperform linear classifiers and domain heuristics while generalizing effectively across test weeks, product departments, and cold-start items.

Future Directions: Four concrete extensions will advance this research: (1) *Promotional Duration ('Weeks-on-Sale') Survival Features*: engineering a counter of consecutive weeks an item has remained at its current discounted price to distinguish 1-week sales from 2-week promotions and reduce False Positives; (2) *Cross-Retailer Entity Resolution*: parsing brand names, package weights, and unit sizes from product titles to match identical products across King Kullen, Stop & Shop, and Gala Fresh and test whether competitor price changes lead local adjustments; (3) *Weekly Flyer NLP & Calendar Features*: integrating TF-IDF + TruncatedSVD embeddings of weekly promotional text and major grocery holiday indicators (Memorial Day, July 4th, Labor Day); and (4) *Multi-Class Directional Forecasting*: extending the binary >=5% threshold into a three-class prediction of price markdowns (<= -5%), price rollbacks/hikes (>= +5%), and price stability.

9. References

[1] F. Valdez, 'KingKullenResearch: Archived longitudinal public catalog snapshots (Freshop platform),' GitHub repository, revision bd93df8e1581, May–Sep. 2026. [Online]. Available: https://github.com/frankstop/KingKullenResearch

[2] F. Valdez, 'StopShopResearch: Archived longitudinal public catalog snapshots,' GitHub repository, revision fe4e9f7579cd, Jul. 2026. [Online]. Available: https://github.com/frankstop/StopShopResearch

[3] F. Valdez, 'GalaFreshBaldwin: Archived public catalog snapshots (SelfPoint platform),' GitHub repository, revision 659adcfbf893, Jul.–Aug. 2026. [Online]. Available: https://github.com/frankstop/GalaFreshBaldwin

[4] F. Valdez, 'GalaFreeport: Archived public catalog snapshots (My Cloud Grocer platform),' GitHub repository, revision 50a4dfff8e46, Jul.–Sep. 2026. [Online]. Available: https://github.com/frankstop/GalaFreeport

[5] Scikit-learn Developers, 'Common pitfalls and recommended practices: Preventing data leakage,' scikit-learn 1.9.1 Documentation, 2026. [Online]. Available: https://scikit-learn.org/stable/common_pitfalls.html#data-leakage

[6] Scikit-learn Developers, 'StandardScaler and SimpleImputer preprocessing API,' scikit-learn 1.9.1 Documentation, 2026. [Online]. Available: https://scikit-learn.org/stable/modules/preprocessing.html

[7] F. Pedregosa et al., 'Scikit-learn: Machine learning in Python,' Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.

[8] L. Breiman, 'Random forests,' Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.

[9] J. H. Friedman, 'Greedy function approximation: A gradient boosting machine,' Annals of Statistics, vol. 29, no. 5, pp. 1189–1232, 2001.

[10] G. Ke et al., 'LightGBM: A highly efficient gradient boosting decision tree,' in Advances in Neural Information Processing Systems (NeurIPS), vol. 30, pp. 3146–3154, 2017.

[11] P. B. Ellickson and S. Misra, 'Supermarket pricing strategies,' Marketing Science, vol. 27, no. 5, pp. 811–828, 2008.

[12] T. Saito and M. Rehmsmeier, 'The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets,' PLOS ONE, vol. 10, no. 3, e0118432, 2015.

[13] W. McKinney, 'Data structures for statistical computing in Python,' in Proc. 9th Python in Science Conf. (SciPy), pp. 56–61, 2010.

[14] C. R. Harris et al., 'Array programming with NumPy,' Nature, vol. 585, pp. 357–362, 2020.