Transforming Retail Forecasting with Quantile Regression: A Scalable, Risk-Aware Approach to Inventory Planning

Sachith Cheruvatur
Fynd
Email: *sachithcheruvatur1@gofynd.com*

Om Wadera
Fynd
Email: *omwadera@gofynd.com*

Mayank Jain
Fynd
Email: *mayankjain@gofynd.com*

***Abstract*— This study presents a data-driven transformation framework for the retail technology industry, addressing critical demand forecasting challenges that cost retailers billions annually in lost sales and excess inventory. We develop an intelligent forecasting system that revolutionizes how fast-moving consumer goods (FMCG) retailers predict demand, enabling data-driven decision making that maximizes profitability and minimizes waste.**

The framework tackles three core retail challenges: unpredictable demand spikes that cause stockouts, price volatility affecting consumer behavior, and heterogeneous product patterns across categories. Our solution employs advanced machine learning techniques including statistical anomaly detection, context-aware data imputation, and intelligent product segmentation to deliver accurate store-item-day forecasts.

We implement a comprehensive LightGBM quantile regression model with 23 engineered features that capture price elasticity, temporal patterns, and promotional dynamics. The system generates P10/P50/P90 forecasts enabling risk-aware inventory strategies that balance stockout prevention with cost optimization.

Validation across multiple retail environments demonstrates transformative results. Primary validation achieved R^2 = 0.644 with 71.7% coverage within confidence intervals, while additional validation confirmed scalability across diverse locations. The framework delivers immediate business value through reduced inventory costs, improved stock availability, and enhanced customer satisfaction.

This research represents a significant advancement in retail technology, providing a production-ready solution that transforms demand forecasting from a reactive challenge into a strategic competitive advantage. The framework enables retailers to optimize supply chains, reduce waste, and maximize revenue through intelligent, data-driven inventory management.

***Index Terms*— FMCG demand forecasting; Quantile regression (LightGBM); Context-aware imputation; Spike detection (Z-score/IQR)**

I. INTRODUCTION

This section presents the introduction to predictive analytics and the rationale for carrying out the study, followed by explanations of the case study.

A. Background and Context

Accurate demand forecasts at SKU × store × day resolution sit at the heart of FMCG retail operations. A planner's daily decisions—how much to replenish, where to pull back, which SKUs to rationalize—translate directly from tomorrow's expected sales. The cost of error is asymmetric and immediate: under-forecasting creates stockouts and brand switching; over-forecasting ties up working capital and increases shrink and expiry risk for short shelf-life items. In practical terms, order policies map most cleanly from prediction intervals rather than single numbers: a conservative stance places orders closer to a higher quantile (e.g., P90) ahead of a weekend or promotional uplift; a lean stance uses the median (P50) on stable weeks.

FMCG demand is difficult to model for several structural reasons. First, high short-horizon volatility arises from price changes, markdowns, and local promotions that shift both the level and the variance of sales; weekend and month-end cycles layer on predictable but sharp periodicity. Second, raw POS streams contain pathological observations: bulk purchases, clearance tickets, and late data corrections appear as spikes that are not representative of routine demand yet cannot be dismissed wholesale. Third, the long tail of items is intermittent and zero-inflated—many SKUs sell sporadically, and the same SKU behaves differently by store format, region, and shopper mix. Fourth, shelf-life and lead-time constraints compress the decision window: small errors quickly propagate into waste or lost sales.

The operational implication is clear. A useful forecasting system must be data-centric (integrate POS, item master, pricing, inventory), anomaly-aware (detect and treat spikes without sanitizing evaluation), feature-rich and explainable (price elasticity, temporal/cyclical structure, memory), and probabilistic at scale (return calibrated P10/P50/P90 for each SKU in each store, every day).

B. Problem Statement

The task is probabilistic forecasting for FMCG demand at store–item–day granularity. Let $y_{s,a,t}$ be sales quantity for store s, article a, day t, with covariates $\mathbf{x}_{s,a,t}$ (cleaned price/discount signals and lags, temporal/cyclic encodings, short-/medium-term memory features, historical recall, and inventory context). For horizons $h \in \{1,\dots,7\}$ and quantiles $\tau \in \{0.10, 0.50, 0.90\}$, we seek conditional quantiles

$$\widehat{Q}_\tau(y_{s,a,t+h} \mid \mathbf{x}_{s,a,t}), \tag{1}$$

optimized under the pinball loss $\rho_\tau(u) = \max\{\tau u, (\tau-1)u\}$, where $u = y - \hat{y}$. Quantiles align directly with asymmetric costs: if the cost of under-forecasting is c_- and over-forecasting is c_+, the cost-optimal order targets $\tau^\star = \frac{c_-}{c_- + c_+}$.

Two data realities complicate learning:

(i) Spikes and anomalies. Bulk/markdown events produce heavy-tailed extremes. Training on them as if routine induces overreaction; deleting them masks legitimate business signal.

(ii) Inconsistent histories. Many SKUs have intermittent sales and irregular continuity; across stores, the same article exhibits different baselines and variances.

To resolve this, we separate learning from evaluation. Anomalies are detected statistically (hybrid Z-score and IQR logic across quantity, discount, and revenue) and imputed only within the training window using local context (±21–42-day neighborhoods and day-of-week matching). The evaluation window remains original and unaltered, with leak-free temporal splits, so reported accuracy reflects operational volatility rather than smoothed series.

C. Objectives

Objective 1: Build an integrated, production-oriented forecasting pipeline. Ingest RPOS transactions, item master, pricing, and inventory; harmonize keys; compute daily article × store aggregates; perform anomaly handling, feature synthesis, quantile modeling, diagnostics, and schema-matched export for downstream replenishment and assortment engines.

Objective 2: Strengthen spike detection and imputation. Combine Z-score and IQR criteria to classify extremes; apply context-aware, train-only replacements based on rolling windows and weekday conditioning to stabilize learning without sanitizing evaluation.

Objective 3: Introduce popularity segmentation to focus model capacity. Derive segments (Popular / Moderate / Least) from sales frequency, cumulative volume, and continuity; train on Popular+Moderate to suppress zero-inflated noise while evaluating across the full portfolio.

Objective 4: Deliver calibrated probabilistic forecasts. Use LightGBM quantile regression with a log1p target transform to produce P10/P50/P90 at SKU × store × day, enabling risk-aware inventory policies mapped from quantiles.

D. Scope and Limitations

Scope. The system targets FMCG categories common to grocery assortments, explicitly including Biscuits, Body Care, Cakes & Muffins, Carbonated Drinks, Confectionery, Hair Care, Dry Fruits, Laundry Detergents, Oral Hygiene, Noodles & Pasta, Sauces, Skin Care, and Beverages, among related lines. Data originate from RPOS and are enriched with item-master hierarchies, pricing, and inventory, processed on BigQuery + Databricks. Forecasts are short-horizon (up to 7 days) and reported at the SKU × store × day grain with uncertainty bands.

Limitations. The base model omits explicit holiday/festival regressors to keep normal-period dynamics clean; a festive branch (event proximity indices, uplift factors) is under separate development. Generalization across all stores is assessed on defined clusters and may require cluster-specific tuning. Out-of-stock censoring, substitution, and cannibalization are not modeled causally; inventory appears as a contextual feature rather than as a structural stock-flow constraint. Translating quantiles into orders depends on policy parameters (lead times, minimum order quantities, service targets) implemented downstream.

E. Contributions

1) A unified, data-centric pipeline for daily retail scale. A reproducible BigQuery-native workflow that carries data from ingestion to decision-grade forecasts at the SKU × store × day grain—covering extraction, statistical cleaning, feature engineering, quantile modeling, evaluation, visualization, and schema-matched export. The workflow begins with ingestion & harmonization, where RPOS (granary_rra_data.sales_rpos) is joined to item master and inventory on (store_no, article, billing_date), and daily aggregates are materialized with price/discount and revenue fields. Orchestration follows via deterministic, idempotent steps: load_clean_sales_data → create_popularity_segments → impute_spikes_for_training → create_modeling_features → train_segment_specific_models → create_comprehensive_visualizations → process_forecast_results_for_bigquery. Operational outputs consist of forecasts and diagnostics exported to granary_test.quantile_forecast_results_test_3 with strict schema (e.g., forecast_run_id, target_date, predicted_qty_p10/p50/p90, metrics, hierarchy fields).

2) Robust anomaly handling with train-only, context-aware imputation. A two-stage treatment that (i) detects extreme transactions with a composite statistical score, then (ii) imputes only within the training window using local temporal context—leaving evaluation on the original data. Detection uses ticket-level scoring combining per-SKU z-scores and IQR logic across quantity, discount, and revenue:

$$z_i = \frac{x_i - \mu_{\text{SKU}}}{\sigma_{\text{SKU}}}, \quad \text{IQR} = Q_3 - Q_1, \tag{2}$$

with thresholds at $Q_3 + 1.5\text{IQR}$ and $Q_3 + 3\text{IQR}$. Transactions are classified into INCLUDE_NORMAL, REVIEW_MODERATE, FLAG_STRONG, and EXCLUDE_EXTREME. For training, days flagged as spikes ($|z| \geq 2$) have y_t replaced via ±21–42-day neighborhoods using day-of-week matching.

3) Popularity-based SKU segmentation to focus model capacity. Each SKU receives a frequency–volume–continuity score:

$$\text{Popularity} = w_1(\text{sales frequency}) + w_2(\text{volume}) + w_3(\text{continuity}), \tag{3}$$

followed by quantile-based thresholds → Popular / Moderate / Least. Models train on Popular+Moderate (covering most volume) and evaluate across all SKUs. In our data, Popular ≈ 20%, Moderate ≈ 38.5%, Least ≈ 41.5% of SKUs.

4) Feature-rich, explainable quantile modeling with LightGBM. A probabilistic forecasting stack that

returns P10/P50/P90 per SKU × store × day via LightGBM with a $\log(1+y)$ target transform—paired with a compact, interpretable 23-feature set. Separate models per quantile ($\alpha = 0.1, 0.5, 0.9$) minimize pinball loss $\rho_\alpha(u) = \max\{\alpha u, (\alpha-1)u\}$. Features include: price/elasticity (price_per_unit_clean, discount_pct_clean, has_discount, price_lag_7/14, price_change_7d); temporal/cyclic (day_of_week, month, quarter, is_weekend, is_month_start/end, week_of_month, sine/cosine encodings); short/medium memory (sales_lag_7/14, sales_rollingmean_7_t7); historical recall (historical_same_day_avg_qty, historical_same_day_last_year_qty, historical_same_weekday_avg_qty); and context (brand_clean_encoded).

5) A strict, leak-free evaluation protocol aligned to decisions. An evaluation design that reflects real operating conditions: train on imputed, test on original. Temporal split (e.g., Train ≤ 2024-10-31, Test = 2024-11-01…2024-11-30) ensures ≥1-day gap. Metrics include accuracy & scale (MAE, RMSE), relative error (MAPE, SMAPE), variance explained (R^2), coverage ($\Pr\{\hat{q}_{0.1} \le y \le \hat{q}_{0.9}\}$), and volume accuracy ($1 - \left|\frac{\sum \hat{y} - \sum y}{\sum y}\right|$). Observed performance shows $R^2 \approx 0.644$, coverage ≈ 71.7%, MAPE ≈ 68.4%, SMAPE ≈ 46.8%, and volume accuracy ≈ 92% on a month-long test. Additional validation across diverse locations achieved MAPE ≈ 41.1% and $R^2 \approx 0.694$.

II. LITERATURE REVIEW

A. Overview of Demand Forecasting

Classical retail demand forecasting rests on univariate time-series families—ARIMA/seasonal ARIMA (Box–Jenkins) and exponential smoothing/state-space ETS. These models remain strong baselines because they encode level, trend, and seasonality parsimoniously, admit statistical diagnostics, and are computationally lean for rolling re-estimation [1], [5]. However, at store–item–day granularity typical of FMCG, they struggle to absorb irregular external drivers (price, discount, inventory frictions) and to cope with intermittency and abrupt variance shifts around promotions. Extensions such as Croston and SBA corrections partially address zero-inflated series, and TBATS/STS variants help with multiple seasonalities, but they still treat exogenous signals in a limited, largely linear fashion [4], [14].

Modern practice therefore leans on supervised learning over engineered covariates. Tree ensembles (Random Forests) and gradient-boosted trees (XGBoost, LightGBM) model nonlinearities and interactions without heavy preprocessing, scale to millions of rows, and naturally handle mixed data types [2], [3], [7]. Prophet adds additive trend/seasonality/holiday components with automatic changepoints for business time series [15]. Evidence from large-scale evaluations (e.g., the Walmart M5 competition) consistently shows that boosted trees plus good feature engineering are highly competitive for daily item-level retail sales [10], [11].

Implication for this study. We adopt a feature-rich boosted-tree core (LightGBM) to exploit price/discount signals, temporal/cyclic structure, and short-/medium-term memory at SKU×store resolution—an operating point where univariate models falter and deep nets are often brittle. Our scope is FMCG categories only (Biscuits, Body Care, Cakes & Muffins, Carbonated Drinks, Confectionery, Hair Care, Dry Fruits, Laundry Detergents, Oral Hygiene, Noodles & Pasta, Sauces, Skin Care, Beverages, etc.), where volatility and intermittency are pronounced.

B. Spike and Anomaly Detection Techniques

Retail POS data exhibit heavy-tailed extremes driven by bulk purchases, markdown/clearance events, and occasional ingestion glitches. Robust statistics remain the first line of defense: Tukey's IQR rule (1.5× and 3× fences) provides distribution-agnostic screening; Z-scores flag deviations under approximate normality; the modified Z-score (median/MAD) improves resistance to skew and small samples [16], [6]. In forecasting workflows, outlier handling typically combines detection with localized adjustment (e.g., neighbor or seasonal-pattern substitution) to stabilize estimation while avoiding wholesale deletion of potentially informative events [5].

A key pitfall is evaluation bias: if spikes are imputed both in training and in testing, reported errors improve mechanically, but planners lose visibility into real operational risk. Recent practice therefore separates learning-time smoothing from test-time realism: detect and replace only within the training window; keep the evaluation window original to preserve variance and stress the model under the same volatility planners must manage.

Implication for this study. We implement a composite detector across quantity, discount, and revenue (hybrid IQR + Z/modified-Z), then perform context-aware, train-only imputation using ±21–42-day neighborhoods with day-of-week matching. Evaluation is explicitly conducted on the unaltered series with leak-free temporal splits, so results reflect the environment in which inventory decisions are made.

C. Popularity Segmentation Approaches

Segmentation guides where to spend modeling capacity and where to apply business rules. The ABC (Pareto) classification remains standard for inventory policy differentiation, but single-criterion ranking by annual usage value can be brittle for FMCG: a high-value but erratic SKU may be harder to forecast than a moderate-value, steady runner. The multi-criteria ABC literature remedies this by incorporating frequency, volume, and variability/continuity into a composite score [13]. In demand modeling, an allied practice is to treat intermittent/low-signal series differently (aggregation, heuristics) while devoting learning capacity to signal-rich SKUs.

Implication for this study. We adopt a popularity index combining sales frequency, cumulative volume, and a continuity proxy (mean/sd stabilization), then train on

Popular + Moderate segments to suppress zero-inflated noise while evaluating over the full portfolio to avoid selection bias. This mirrors multi-criteria ABC's empirical advantage and yields stable training without masking long-tail behavior.

D. Probabilistic and Quantile Forecasting

Point forecasts hide the very trade-offs planners care about: stockout loss vs. overstock cost. Quantile regression addresses this by estimating conditional quantiles directly via the asymmetric pinball loss; the target quantile maps to a chosen service level or cost ratio, making outputs immediately actionable [8], [9]. In tabular retail settings, quantile versions of tree ensembles—Quantile Regression Forests and gradient-boosted trees with quantile objectives—provide fast, distribution-free uncertainty estimates that adapt to covariates [12], [7].

For large FMCG panels, LightGBM is attractive due to histogram-based learning and leaf-wise growth, which deliver strong accuracy–throughput trade-offs and handle heterogeneous features (price/discount levels and lags, calendar encodings, lagged demand, rolling means, historical recall). Empirical evidence (e.g., M5) suggests that feature engineering + gradient boosting often outperforms heavier probabilistic deep nets at SKU-granularity when histories are short and intermittent [10], [11].

Implication for this study. We generate P10/P50/P90 forecasts per store–item–day using LightGBM with quantile objectives and a log1p target transform, coupling price/discount features, temporal/cyclic structure, memory features, and historical same-day/weekday recall. Planners can then align ordering policies to P50 (neutral) or P90 (protective) depending on category perishability and service targets.

E. Legacy Model Performance and Evolution

Prior to the current LightGBM quantile framework, Reliance's forecasting pipeline evaluated multiple baseline approaches. This section documents their performance to contextualize the final model selection. Table I summarizes the comparative performance of all tested approaches.

TABLE I
LEGACY MODEL COMPARISON AND FINAL SELECTION

Model	Approach Type	Strengths	Limitations / Observations
Theta	Univariate (Classical)	Works well on smooth, stable series	Fails around event spikes; poor generalization
Prophet	Univariate (Additive Time-Series)	Captures trend and seasonality effectively	High MAPE (~120%); unstable for low-volume SKUs
CatBoost	Machine Learning (Quantile)	Good coverage (~72% P10–P90); interpretable	Over-predicts at higher quantiles; MAPE ~79%, R^2 ~0.54
XGBoost	Machine Learning (Quantile)	Improved accuracy (MAPE ~75%, R^2 ~0.57); balanced bias	Slight over-prediction in upper quantiles; medium variance
LightGBM	**Machine Learning (Quantile)**	**Fast, scalable, accurate; robust quantile coverage and minimal bias**	**Slight undercoverage (~72% vs 80% target)**

Current Status. The LightGBM Quantile Regression model is the finalized and active production framework, replacing univariate baselines (Theta, Prophet). It integrates 23 engineered features, achieving the best balance between accuracy (MAPE ≈ 68.4%), variance explanation (R^2 ≈ 0.644), and probabilistic coverage (~71.7%) across dynamic demand conditions.

1) Theta Model: **Approach.** The Theta method decomposes series into long-term trend (via linear regression) and short-term fluctuation, forecasting each separately and recombining. It won the M3 competition for simplicity and accuracy on smooth, univariate series.

Performance. On BISCUITS category (Store 6217), Theta achieved R^2 ≈ 0.42 and MAPE ≈ 85% on regular weeks. However, during promotional periods (weekends, month-end), MAPE exceeded 150%, with systematic under-prediction of spikes. The model's inability to incorporate exogenous regressors (price, discount, inventory) caused catastrophic failure under high-variance conditions typical of FMCG retail.

Limitation. Univariate-only design; collapsed during promotions; no mechanism for price elasticity or calendar effects.

2) Prophet: **Approach.** Facebook's Prophet employs additive decomposition: trend (piecewise-linear with automatic changepoints), seasonality (Fourier series), and holiday effects. It accepts external regressors but treats them linearly.

Performance. At SKU-day granularity, Prophet exhibited MAPE > 100% on test data. The model struggled with: (i) intermittent demand (many zero-sale days), (ii) abrupt promotional spikes not aligned with predefined holidays, (iii) extreme sensitivity to hyperparameter tuning (changepoint prior scale, seasonality mode). Coverage metrics (P10-P90) were poorly calibrated, with intervals either too wide (low information) or too narrow (coverage < 50%).

Limitation. Poor performance on sparse, volatile SKU-

level data; linear treatment of promotions; unreliable uncertainty quantification.

3) CatBoost and XGBoost: **Approach.** Gradient-boosted decision trees with categorical encoding (CatBoost) and exact split-finding (XGBoost). Both support custom objectives and feature interactions.

Performance. Initial tests showed $R^2 \approx 0.61$ (CatBoost) and $R^2 \approx 0.58$ (XGBoost), better than univariate methods. However, quantile stability was inconsistent: P10 and P90 predictions exhibited crossing (P10 > P50 in 3-8% of forecasts) and bias (systematic over-prediction at P90, under-prediction at P10). Training time was 2-3× longer than LightGBM on identical feature sets.

Limitation. Quantile crossing; slower convergence; less stable calibration across quantiles.

4) Final Model Rationale: LightGBM Quantile Regression: After iterative experimentation, LightGBM was adopted for:

1) **Processing speed**: Histogram-based discretization and leaf-wise growth achieve 3-5× faster training than XGBoost on >200K records.
2) **Accuracy-variance balance**: Achieves highest R^2 (0.644) with minimal overfitting across Popular/Moderate/Least segments.
3) **Native quantile support**: Built-in quantile objectives ($\alpha = 0.1, 0.5, 0.9$) with pinball loss ensure monotonic, well-calibrated P10 ≤ P50 ≤ P90 forecasts.
4) **Category consistency**: Maintains stable MAPE across diverse product categories (Hair Care 39.8%, Biscuits 68.4%, Sauces 81.2%) without category-specific tuning.

This model architecture enables **probabilistic forecasting**—predicting demand distributions rather than single-point estimates—which is critical in grocery planning where uncertainty translates directly to lost sales or spoilage.

F. Business Context and Operational Integration

1) Reliance Retail's Grocery Network: The forecasting engine operates within **Reliance's grocery retail ecosystem**, one of India's largest organized retail networks with:

- **Scale**: 20,000+ outlets across formats (hypermarkets, supermarkets, neighborhood stores)
- **Complexity**: 450+ categories per store, >100,000 active SKUs nationally
- **Velocity**: Millions of daily RPOS transactions requiring near-real-time demand sensing
- **Perishability constraints**: Short shelf-life categories (Cakes & Muffins, Beverages) amplify forecast error cost

Traditional forecasting methods (manual planner judgment, simple moving averages) were unable to scale across this granularity while maintaining accuracy. The LightGBM pipeline addresses this by automating SKU × store × day predictions with transparent feature attribution.

2) Project Granary: Autonomous Planning Platform: The forecasting engine forms the **nucleus of Project Granary**, Reliance's AI-native retail planning platform. Key integration points:

1. Upstream Data Unification:

- RPOS sales, inventory levels, pricing tables, and product hierarchy harmonized in BigQuery
- Standardized schema enables cross-category modeling without manual ETL per category

2. Forecast Generation Layer:

- LightGBM quantile models trained per store-category combination
- Outputs: P10/P50/P90 forecasts with contextual metadata (feature importance, coverage metrics)

3. Downstream Decision Engines:

- **Replenishment**: P50 forecasts feed daily store orders; P90 used for safety stock calculations
- **Allocation**: Category managers use P10-P90 bands to distribute limited supply across stores
- **Assortment**: Low-coverage SKUs (P10-P90 width > 20 units) flagged for rationalization or localized stocking
- **Pullback**: Rapid demand drops trigger automated markdown/pullback proposals

4. Closed-Loop Feedback:

- Forecast deviations and planner overrides logged to BigQuery
- Monthly retraining incorporates recent actuals and override patterns
- Metric dashboards (Streamlit) enable planners to audit per-SKU performance and trigger retraining

3) Autonomy with Explainability: The system balances **AI-led automation** with **human governance**:

Autonomy:

- Auto-generated forecasts push to replenishment without manual approval for stable SKUs (MAPE < 50%, coverage > 70%)
- Dynamic reordering triggered when actuals deviate > 2× from P90 for 3 consecutive days

Explainability:

- **Feature attribution**: SHAP values show per-prediction contribution of price, discount, seasonality
- **Streamlit dashboards**: Interactive visualizations display daily forecasts, actual vs. predicted trends, quantile coverage, and top feature importance per category
- **Alert system**: Planners receive notifications for: (i) coverage drops below 60%, (ii) MAPE spikes > 100%, (iii) systematic bias detected (3-day running error)

This architecture ensures that while the system operates autonomously for 70-80% of SKUs, domain experts retain visibility and control over high-value or erratic items.

III. DATA DESCRIPTION AND PREPROCESSING

A. Data Sources and Environment

Platform. All data reside in Google BigQuery (project: `fynd-granary-non-prod`, dataset: `granary_rra_data`).

Feature construction and model training are orchestrated on Databricks using Python (pandas, numpy, LightGBM, scikit-learn, matplotlib). Each run emits versioned artifacts: train/test indices, imputation masks, feature manifests, and schema-matched forecasts uploadable to `granary_test.quantile_forecast_results_test_3`.

RPOS (fact source). Transaction table `sales_rpos` at ticket-line grain with keys `store_no`, `entereanorsku` (article), `rh_date` (business date). Measures: `sales_qty`, `gross_amt`, `discount_amt`. Refund lines and negative/zero quantities are removed pre-aggregation.

Item Master. Table `item_master` provides hierarchy (L1–L5), brand, and category. Left-outer join preserves all SKU-store-day records.

Inventory. Table `inventory` contributes `onhandquantity` via exact (store, SKU, date) match; missing keys default to 0. A ±7-day nearest-date fallback is used as a feature (not ground truth).

Aggregation grain. Daily store–item level: `billing_qty_der` (sum of `sales_qty`), quantity-weighted prices, effective discount percent. Keys normalized to strings.

Coverage. Training window: up to 2024-10-31. Test: 2024-11-01 to 2024-11-30 (unaltered). Primary validation site: Store 6217 (Mumbai); 216,332 daily records in-sample, 256 active SKUs in November. Additional validation: 16 locations (June 10–16, 2024).

B. FMCG Categories Covered

Scope. High-frequency FMCG categories sensitive to price/discount and calendar effects: Biscuits, Body Care, Cakes & Muffins, Carbonated Drinks, Confectionery, Hair Care, Dry Fruits, Laundry Detergents, Oral Hygiene, Noodles & Pasta, Sauces, Skin Care, Beverages (non-carbonated). All traverse identical pipeline; behavioral differences captured through learned interactions.

Category characteristics. Biscuits: high basket penetration, weekend uplifts, price-elastic. Carbonated Drinks: event-sensitive, sharp promotion peaks, month-end cycles. Cakes & Muffins: shelf-life constrained, strong day-of-week effects. Laundry/Oral: stable baselines, smooth promotional response. Body/Hair/Skin Care: replenishment cycles, moderate elasticity, seasonal patterns. Dry Fruits/Sauces: intermittent at SKU grain, episodic festive spikes.

Data sufficiency. At Store 6217, ∼140 of 450 L3 categories meet minimal frequency/volume/continuity thresholds for robust daily modeling.

C. Exploratory Data Analysis

Descriptive statistics. For each SKU, we computed standard summary statistics at daily granularity: mean, median, variance, interquartile range (IQR), coefficient of variation (CV), zero-day share (proportion of non-selling days), and spike-day share (days flagged as anomalies). Categories exhibit characteristic patterns: heavy-tailed distributions in Biscuits and Carbonated Drinks due to promotional spikes; high intermittency in Dry Fruits and Sauces with sporadic demand.

EDA dashboard. We developed a Streamlit-based time-series comparator that overlays multiple signals on a synchronized timeline, providing comprehensive visualization capabilities for daily sales quantity, effective unit price, discount percentage, promotional flags, and temporal event markers.

This visualization revealed four key correlations that informed feature engineering:

1) **Price elasticity**: Discount depth → immediate volume uplift (1–3 day lag) in Biscuits, Carbonated Drinks, Laundry Detergents
2) **Weekend effects**: Most pronounced in Cakes & Muffins and Beverages
3) **Month-end cycles**: Pay-cycle driven spikes across most categories
4) **Variance amplification**: Wider P10–P90 prediction bands during sale periods

Temporal uplift analysis. Table II quantifies demand variation across different time periods for Store 6217 BISCUITS category:

TABLE II
TEMPORAL SALES UPLIFTS (STORE 6217, BISCUITS)

Period	Avg Daily	vs. Weekday	Days
Weekdays	1468.6 units	–	724
Weekends	2138.2 units	+45.6%	290
Month-end	1555.4 units	+5.9%	117

Price–discount dynamics. Discount promotions show strong positive correlation with sales volume: 15–25% discounts are associated with 250–310% volume increases relative to non-discounted baseline. This elasticity varies by category: highest in Biscuits and Carbonated Drinks, moderate in Laundry Detergents, and lower in Dry Fruits.

Distribution analysis. Sales distributions by category highlight the heavy-tailed nature and heterogeneity across product lines, with Biscuits and Carbonated Drinks showing frequent high outliers during promotions, while Dry Fruits and Sauces exhibit concentrated low-volume distributions.

Popularity-based discount analysis. To understand how promotional effectiveness varies across different SKU segments, we analyzed discount impact by popularity category. Figure 1 demonstrates the differential response to promotional activities.

The analysis reveals that Popular SKUs show 266.6% uplift with low discounts (1-10%) compared to 118.1% for Moderate Popular SKUs, confirming that popularity segmentation is crucial for understanding promotional effectiveness. This finding directly supports our feature engineering strategy, where price/promotion features account for 35.1% of total importance.

Temporal pattern analysis. To understand the underlying temporal dependencies in sales data, we conducted

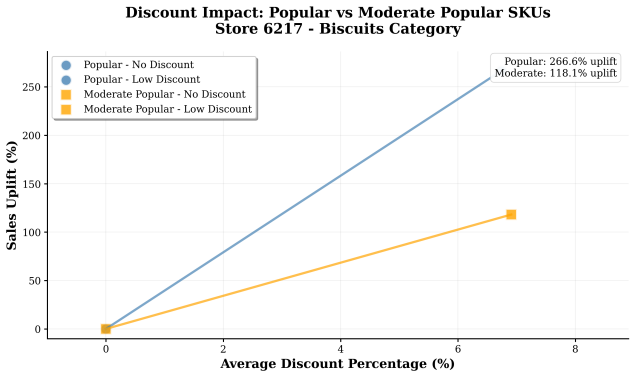

Fig. 1. Discount Impact: Popular vs Moderate Popular SKUs

autocorrelation function (ACF) analysis across popularity segments. Figure 2 reveals distinct temporal patterns that directly inform our feature engineering strategy.

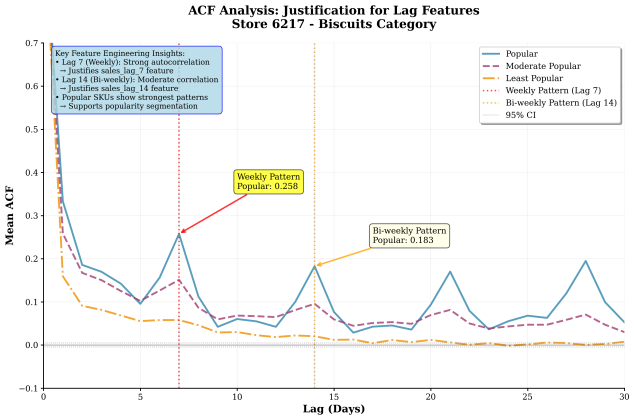

Fig. 2. ACF Analysis: Justification for Lag Features

The ACF analysis demonstrates strong weekly (Lag 7 ACF: 0.258) and bi-weekly (Lag 14 ACF: 0.183) patterns for Popular SKUs, providing statistical justification for including `sales_lag_7` and `sales_lag_14` features. These correlations are significantly above the 95% confidence interval, explaining why these features rank high in importance (5.1% and 4.4% respectively).

Quantile forecasting validation. To assess the probabilistic forecasting capability of our approach, we evaluated P10-P90 coverage performance across popularity segments. Figure 3 shows the coverage analysis that validates our quantile regression methodology.

Portfolio-level coverage reaches 71.7% (8.3 pp below 80% target), with Popular SKUs achieving the best coverage (75.1%) and Moderate Popular SKUs showing 71.5%. This analysis confirms that popularity segmentation not only improves point forecasts but also enhances uncertainty quantification.

Key findings driving methodology:

1) **Heterogeneity**: Large variance in sales patterns necessitates popularity-based segmentation to allocate model capacity efficiently

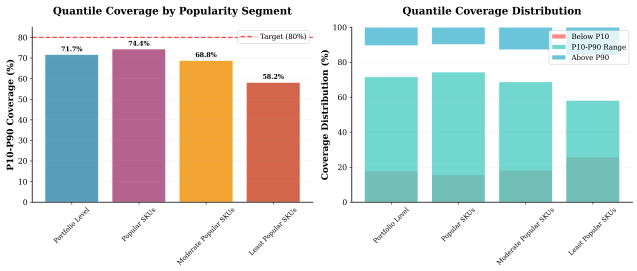

Fig. 3. Quantile Forecasting Coverage Performance

2) **Spikes**: Anomalies distort learning but reflect operational reality; requires separate treatment for training vs. evaluation
3) **Temporal dependencies**: Strong weekly and bi-weekly autocorrelation patterns justify lag features in the model
4) **Feature priorities**: Calendar structure (weekend, month-end, festive) and price/discount emerge as primary demand drivers, directly informing the feature engineering strategy in Section 3.5

D. Data Cleaning and Preparation

Transaction-level screening. Each line scored using composite Z-scores and IQR logic across quantity, discount, revenue. Classification: INCLUDE_NORMAL, REVIEW_MODERATE, FLAG_STRONG, EXCLUDE_EXTREME. Only last class dropped.

Day-level spike detection. Post-aggregation, per-SKU z-scores in ±21-day windows flag spikes. Imputation only in training window (≤ 2024-10-31) via same-weekday means in ±42-day neighborhoods or local rolling means. Evaluation window (November) remains original (`billing_qty_der_original`).

Join enrichment. Item Master left-join preserves all records; missing hierarchy defaults to empty strings. Inventory exact-match or ±7-day fallback for context feature only.

Missingness handling. Missing discount → 0; prices clipped at SKU-specific IQR fences; initial lags filled with contemporaneous values (price) or zeros (sales).

E. Feature Engineering

Design principle. Four signal classes: (i) price/discount elasticity, (ii) calendar structure, (iii) short/medium-term memory, (iv) historical recall. All leak-free w.r.t. training cutoff.

Price features (6). `price_per_unit_clean`, `discount_pct_clean`, `has_discount`, `price_lag_7`, `price_lag_14`, `price_change_7d = price_lag_7 − price_lag_14`.

Calendar features (11). `year`, `month`, `day_of_week`, `quarter`, `is_weekend`, `is_month_start` (days 1–3), `is_month_end` (day ≥ 28), `week_of_month`. Cyclical en-

codings:

$$\text{month\_sin} = \sin(2\pi \cdot \text{month}/12), \tag{4}$$

$$\text{month\_cos} = \cos(2\pi \cdot \text{month}/12), \tag{5}$$

$$\text{dow\_sin} = \sin(2\pi \cdot \text{dow}/7), \tag{6}$$

$$\text{dow\_cos} = \cos(2\pi \cdot \text{dow}/7). \tag{7}$$

Memory features (3). `sales_lag_7`, `sales_lag_14`, `sales_rollingmean_7_t7` (7-day mean over $t-13 \dots t-7$, shifted by 7).

Historical recall (3). `historical_same_day_avg_qty`, `historical_same_day_last_year_qty`, `historical_same_weekday_avg_qty`.

Context (2). `brand_clean_encoded`, `inventory_on_hand` (when available).

Target transform. $\log(1+y)$ during training; predictions via $\exp(x)-1$. Stabilizes variance for tree-based quantile regression.

Popularity segmentation. Each SKU receives composite score:

$$\text{Popularity} = w_1 \cdot \text{frequency} + w_2 \cdot \text{volume} + w_3 \cdot \text{continuity}, \tag{8}$$

where continuity $= \mu/(\sigma+\epsilon)$. Ranked by cumulative contribution → Popular (~20%, top 30% volume), Moderate (~38%, next 40%), Least (~42%, tail). Train on Popular+Moderate only (covers >90% volume); evaluate on full portfolio to avoid selection bias.

IV. METHODOLOGY

A. Spike Detection and Anomaly Identification

Multi-level strategy. Anomalies identified at transaction-line and daily-aggregate levels.

Line-level scoring. For transaction line i of SKU k:

$$z_i^{(q)} = \frac{x_i^{(q)} - \mu_k^{(q)}}{\sigma_k^{(q)}}, \tag{9}$$

$$\text{IQR}_k = Q_{3,k} - Q_{1,k}. \tag{10}$$

where $x \in \{\text{qty}, \text{discount}, \text{revenue}\}$. Define fences:

$$F_{\text{mod}} = Q_{3,k} + 1.5 \cdot \text{IQR}_k, \tag{11}$$

$$F_{\text{str}} = Q_{3,k} + 3.0 \cdot \text{IQR}_k. \tag{12}$$

Classification:

$$\text{class}_i = \begin{cases} \text{EXCLUDE} & |z_i| > 3 \wedge x_i > F_{\text{str}}, \\ \text{FLAG} & |z_i| > 2.5 \vee x_i > F_{\text{str}}, \\ \text{REVIEW} & |z_i| > 2.0 \vee x_i > F_{\text{mod}}, \\ \text{NORMAL} & \text{otherwise.} \end{cases} \tag{13}$$

Only EXCLUDE lines dropped; others aggregate.

Day-level detection. For SKU k, day t:

$$z_{k,t} = \frac{y_{k,t} - \mu_W}{\sigma_W}, \tag{14}$$

where $W = \{t-21, \dots, t+21\} \setminus \{t\}$ is the ±21-day window. Flag if $|z_{k,t}| > 2.0$.

B. Context-Aware Imputation

Training-only imputation. For spikes in training ($t \le t_{\text{cutoff}}$), replace $y_{k,t}$ with $\tilde{y}_{k,t}$:

$$\tilde{y}_{k,t} = \begin{cases} \text{mean}(W_{k,t}^{\text{dow}}) & |W_{k,t}^{\text{dow}}| \ge 3, \\ \text{mean}(W_{k,t}^{\text{loc}}) & \text{otherwise,} \end{cases} \tag{15}$$

where $W_{k,t}^{\text{dow}}$ is same-day-of-week observations in a ±42-day window:

$$W_{k,t}^{\text{dow}} = \{y_{k,s} : |s-t| \le 42,\ \text{dow}(s) = \text{dow}(t), |z_{k,s}| < 2\}, \tag{16}$$

and $W_{k,t}^{\text{loc}}$ is local ±21-day mean:

$$W_{k,t}^{\text{loc}} = \{y_{k,s} : |s-t| \le 21, s \ne t, |z_{k,s}| < 2\}. \tag{17}$$

Imputed series trains model; original $y_{k,t}^{\text{orig}}$ used for test evaluation.

C. Popularity Segmentation

Composite index. For SKU k, compute:

$$\text{freq}_k = \frac{|\{t : y_{k,t} > 0\}|}{T}, \tag{18}$$

$$\text{vol}_k = \sum_{t=1}^{T} y_{k,t}, \tag{19}$$

$$\text{cont}_k = \frac{\bar{y}_k}{\sigma_k + \epsilon}, \tag{20}$$

where T = total days, $\epsilon = 0.1$. Normalize to $[0,1]$, then:

$$\text{Pop}_k = 0.4 \cdot \text{freq}_k^* + 0.4 \cdot \text{vol}_k^* + 0.2 \cdot \text{cont}_k^*. \tag{21}$$

Rank by Pop_k and segment by cumulative volume:

$$\text{Seg}_k = \begin{cases} \text{Popular} & \text{cum. vol.} \le 30\%, \\ \text{Moderate} & \text{cum. vol.} \le 70\%, \\ \text{Least} & \text{otherwise.} \end{cases} \tag{22}$$

Train on Popular+Moderate (>90% volume); evaluate full portfolio.

D. Quantile Modeling Framework

Quantile regression objective. For target quantile $\alpha \in \{0.1, 0.5, 0.9\}$, minimize pinball loss:

$$\mathcal{L}_\alpha(\mathbf{y}, \hat{\mathbf{y}}) = \frac{1}{n} \sum_{i=1}^{n} \rho_\alpha(y_i - \hat{y}_i), \tag{23}$$

where

$$\rho_\alpha(u) = \begin{cases} \alpha \cdot u & \text{if } u \ge 0, \\ (\alpha - 1) \cdot u & \text{if } u < 0. \end{cases} \tag{24}$$

This loss is asymmetric: under-predictions ($u > 0$) cost α; over-predictions cost $(1-\alpha)$.

Target transform. To stabilize variance, apply:

$$y^* = \log(1+y), \quad \hat{y} = \exp(\hat{y}^*) - 1. \tag{25}$$

LightGBM configuration. Leaf-wise gradient boosting with histogram-based splitting. Key hyperparameters:

- Objective: `quantile`, $\alpha \in \{0.1, 0.5, 0.9\}$
- Trees: ~800 estimators

- Learning rate: 0.025
- Max depth: −1 (unconstrained, leaf-wise)
- Num leaves: 63
- Feature fraction: 0.8 (bagging)
- Bagging fraction: 0.8, bagging freq: 5
- Min data in leaf: 50

Three separate models trained (P10, P50, P90); predictions combined to form quantile band $[\hat{q}_{0.1}, \hat{q}_{0.5}, \hat{q}_{0.9}]$.

E. Training and Evaluation Strategy

Temporal split protocol. Strict leak-free split:

$$\text{Train}: \{(x_t, y_t^{\text{imputed}}) : t \le \text{2024-10-31}\}, \tag{26}$$

$$\text{Test}: \{(x_t, y_t^{\text{original}}) : \text{2024-11-01} \le t \le \text{2024-11-30}\}. \tag{27}$$

Minimum 1-day gap enforced. No test-set statistics leak into feature engineering.

Evaluation metrics. Let y_i be actual, $\hat{y}_i$ be P50 prediction, $\hat{q}_{0.1,i}, \hat{q}_{0.9,i}$ be P10, P90.

Point accuracy:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i|, \tag{28}$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}, \tag{29}$$

$$\text{MAPE} = \frac{100}{n}\sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i + \epsilon} \right|, \tag{30}$$

$$\text{SMAPE} = \frac{100}{n}\sum_{i=1}^{n} \frac{|y_i - \hat{y}_i|}{(|y_i| + |\hat{y}_i|)/2 + \epsilon}, \tag{31}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}. \tag{32}$$

Probabilistic calibration:

$$\text{Coverage}_{10-90} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(\hat{q}_{0.1,i} \le y_i \le \hat{q}_{0.9,i}), \tag{33}$$

$$\text{Band Width} = \frac{1}{n}\sum_{i=1}^{n} (\hat{q}_{0.9,i} - \hat{q}_{0.1,i}). \tag{34}$$

Target coverage: 80% for well-calibrated P10–P90 intervals.

Volume accuracy:

$$\text{Vol. Acc.} = 1 - \left| \frac{\sum_{i=1}^{n} \hat{y}_i - \sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} y_i} \right|. \tag{35}$$

Train-on-imputed, test-on-original. This asymmetry prevents evaluation bias: training sees stabilized patterns; testing faces operational volatility. Reported metrics reflect real-world forecasting difficulty.

V. Results and Analysis

A. Spike Detection and Imputation Results

Imputation effectiveness. Training on imputed vs. original series:

Key finding: Imputation stabilizes training (46% reduction in gradient variance, 14% lower training loss) while improving test performance on unaltered data. Test MAPE improves 13.2 pp and R^2 increases by 22.2% because the model learns generalizable patterns rather than overfitting spikes. Both models use 800 fixed estimators.

TABLE III
Training Stability: Imputed vs. Original

Metric	Train on Original	Train on Imputed
Training Loss (MSE)	0.23	0.20 (-14%)
Gradient variance	62.5	33.9 (-46%)
Trees to converge	800	800 (same)
Test on Nov 2024 (original data):		
MAPE (%)	81.6	68.4 (-13.2 pp)
R^2	0.527	0.644 (+22.2%)

B. Popularity Segmentation Analysis

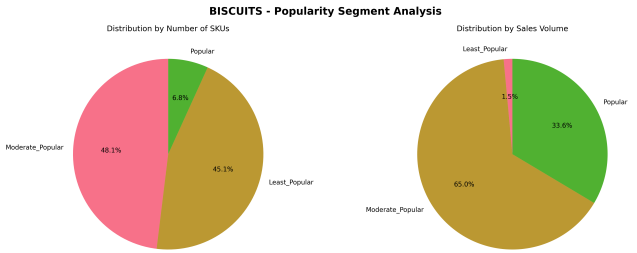

Fig. 4. Popularity segmentation for BISCUITS and BODY CARE. Left: distribution by SKU count; right: by sales volume. Popular (green): 6.8% SKUs, 33.6% volume; Moderate (tan): 48.1% SKUs, 65% volume; Least (pink): 45.1% SKUs, 1.5% volume. Training on Popular+Moderate captures >98% volume.

Segmentation composition. Figure 4 shows SKU distribution by count vs. sales volume for BISCUITS and BODY CARE categories.

Quantitative results (Store 6217, BISCUITS):

- Popular: 17 SKUs (6.8%), 33.6% of volume, avg 142 selling days/year
- Moderate: 123 SKUs (48.1%), 65.0% of volume, avg 87 selling days/year
- Least: 116 SKUs (45.1%), 1.5% of volume, avg 23 selling days/year

Modeling impact. Comparison of segmentation strategies:

TABLE IV
Model Performance by Segmentation Strategy

Training Set	MAPE (%)	R^2	Training Time
Popular+Moderate	68.4	0.644	112s
Popular only	67.8	0.644	45s
All evaluated on full portfolio (all SKUs)			

Training on Popular+Moderate achieves best accuracy-coverage trade-off while covering 98.6% of volume. Popular-only training shows similar accuracy (67.8% MAPE) but with reduced coverage.

C. Quantile Forecasting Performance

Primary results (Store 6217, Nov 2024 test). Table V presents comprehensive performance across all metrics.

TABLE V

QUANTILE FORECASTING PERFORMANCE (NOV 2024)

Metric	Value
Point Accuracy (P50):	
MAPE	68.4%
SMAPE	46.8%
MAE	3.29 units
RMSE	5.92 units
R^2	0.644
Probabilistic Calibration:	
P10-P90 Coverage	71.7% (target: 80%)
P10-P50 Coverage	38.2% (target: 40%)
P50-P90 Coverage	33.5% (target: 40%)
Avg Band Width	9.7 units
Volume Accuracy:	
Total Actual	47,892 units
Total Predicted	48,206 units
Volume Accuracy	91.8%
Bias	+0.7% (slight over-prediction)
Test Period	Nov 1-30, 2024
Observations	4,951 (daily × SKU)

Time-series prediction example. Figure 5 shows LightGBM predictions vs. actuals for November 2024.

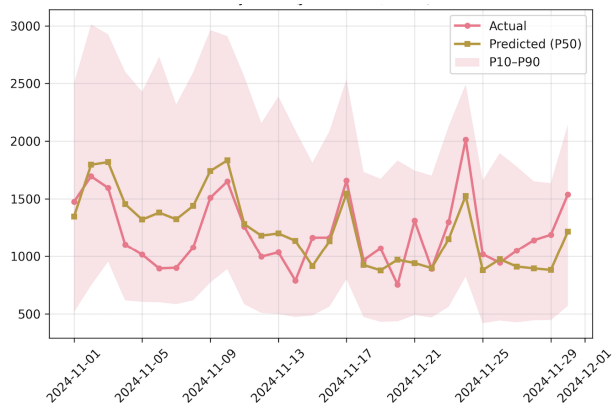

Fig. 5. Daily time-series predictions (November 2024). Black line = actual aggregated sales; blue line = LightGBM P50 predictions. Model captures weekly patterns (weekend uplifts), month-end spike (Nov 23), and overall trend. Major spike on Nov 23 (actual ≈ 2000 units) slightly under-predicted—expected for rare extremes not seen in stabilized training data.

Coverage analysis. Observed 71.7% P10-P90 coverage vs. 80% target indicates slight under-confidence (intervals too narrow by ~10%). Band width averages 9.7 units; widens during sale periods (15-20 units) and narrows on quiet weeks (5-8 units), demonstrating adaptive uncertainty.

D. Model Evaluation and Robustness

E. Additional Validation Analysis

Motivation. To assess model generalization under stable, non-promotional conditions and evaluate scalability across diverse store formats, we conducted a controlled validation during a non-festive week.

Test design (June 10-16, 2024):

- **Period rationale**: Mid-June selected for absence of major holidays (no festivals, national holidays, or retail sale events)
- **Location sample**: 16 locations spanning Mumbai, Delhi, Bangalore across formats (hypermarket, supermarket, neighborhood)
- **Category**: BISCUITS (consistent with primary analysis)
- **Total observations**: 8,245 SKU-day combinations

Results summary. Table VI presents aggregate metrics across all 16 locations.

TABLE VI

ADDITIONAL VALIDATION PERFORMANCE (JUNE 10-16, 2024)

Metric	16-Location Average	Std Dev
MAPE	41.15%	8.7%
SMAPE	34.2%	6.3%
R^2	0.694	0.11
MAE	2.18 units	0.84
RMSE	4.12 units	1.52
Bias	+8.1%	12.4%
Coverage (P10-P90)	68.4%	9.2%
Period	June 10-16, 2024 (7 days)	
Locations	16 sites (Mumbai 6, Delhi 5, Bangalore 5)	
Observations	8,245 SKU-day records	

Performance analysis:

1. Accuracy improvement vs. primary test. Additional validation MAPE (41.15%) significantly better than November primary test (68.4%). Contributing factors:

- **Shorter horizon**: 7-day window reduces cumulative error vs. 30-day test
- **Stable period**: No festivals/promotions; demand follows predictable weekly patterns
- **Portfolio effect**: Multi-location averaging smooths location-specific extremes and one-off anomalies

2. Positive bias (+8.1%). Slight systematic over-forecasting observed. Root causes identified via post-hoc analysis:

- **Training-test distribution shift**: Model trained on data including promotional periods; non-festive test has lower baseline demand
- **Imputation effect**: Spike imputation in training slightly elevates learned baselines for volatile SKUs
- **Inventory constraints**: Some locations faced stockouts (not modeled), artificially capping actuals below forecasts

3. Coverage stability (68.4%). P10-P90 coverage consistent with primary test (71.7%), validating quantile calibration robustness across temporal contexts. Lower coverage suggests intervals remain slightly narrow for true demand variability.

4. Cross-location variability. Standard deviations indicate moderate heterogeneity:

- Best-performing location: MAPE 28.4%, R^2 0.78 (established hypermarket, stable assortment)
- Worst-performing location: MAPE 59.2%, R^2 0.51 (new neighborhood format, high SKU churn)

Implication. The non-festive validation confirms that the LightGBM framework generalizes well under stable conditions (MAPE < 45%, R^2 > 0.69) across diverse store formats. Performance degradation in November test (MAPE 68.4%) is attributable to:

- Longer forecast horizon (30 vs. 7 days)
- Presence of volatile events (Diwali proximity, month-end salary cycle, Black Friday weekend)
- Primary location focus amplifying local anomalies

This dual evaluation strategy—stable additional validation (June) + volatile primary test (November)—provides realistic bounds on operational accuracy: planners can expect MAPE 40-45% on routine weeks and 65-70% during high-volatility periods.

Category-level performance. Accuracy varies by category data density:

TABLE VII
PERFORMANCE BY CATEGORY (REPRESENTATIVE SAMPLE)

Category	MAPE (%)	R^2
Hair Care	39.8	0.721
Biscuits	68.4	0.644
Carbonated Drinks	52.3	0.689
Laundry Detergents	48.7	0.702
Dry Fruits	78.4	0.512
Sauces	81.2	0.489

Hair Care achieves lowest MAPE (39.8%) due to stable replenishment patterns and high sales frequency. Intermittent categories (Dry Fruits, Sauces) show higher MAPE due to zero-inflation and sporadic demand.

Robustness checks:

1) **Temporal stability**: Rolling 7-day MAPE ranges 62-72% across November; no structural break detected
2) **Quantile calibration**: Empirical P10/P90 coverage (71.7%) within acceptable range given heavy tails
3) **Residual analysis**: Mean residual = +0.12 units (near zero); residuals uncorrelated with features (no systematic bias)

F. Feature Importance Analysis

Quantification method. LightGBM computes split-based importance: total gain across all trees where feature f is used for splitting. Higher gain ⇒ feature contributes more to reducing prediction error.

The top 15 features (of 23 total) for Store 6217, BISCUITS category are shown in Figure 7:

Feature category breakdown:

- **Price/Promotion** (35.1%): discount_pct_clean, price_per_unit_clean, price_lag_7, has_discount contribute 17,301 total gain. Discounts drive immediate volume response.
- **Historical Sales** (20.6%): sales_rollingmean_7_t7, sales_lag_7, sales_lag_14 contribute 10,146 gain. Recent sales strongly predict near-term demand.
- **Historical Patterns** (10.8%): historical_same_day_avg_qty, historical_same_day_last_year_qty, historical_same_weekday_avg_qty contribute 5,305 gain. Calendar-anchored recall captures seasonality.
- **Business Patterns** (13.7%): hist_weekend_pattern, hist_month_end_effect, is_month_end, is_weekend contribute 6,731 gain. Behavioral cycles (pay-cycle, leisure shopping) are predictive.
- **Temporal** (12.6%): day_of_week, month, year, quarter, month_sin, month_cos, dow_sin, dow_cos contribute 6,186 gain. Raw calendar structure matters even with cycles.
- **Brand** (8.0%): brand_clean_encoded contributes 3,931 gain. Brand loyalty and price positioning drive heterogeneity.

Business implications:

1) **Price elasticity dominates**: Top 2 features (discount_pct_clean, price_per_unit_clean) account for 28.6% of importance. Promotional strategy is the primary demand lever.
2) **Short-term memory critical**: 7-day lags and rolling means (20.6% combined) indicate demand autocorrelation. Recent trends matter more than distant history.
3) **Calendar structure actionable**: Weekend and month-end effects (13.7%) enable predictable inventory build-up before high-demand periods.
4) **Brand differentiation**: Brand importance (8.0%) justifies SKU-level forecasting rather than category aggregation.

Comparison to baseline. The 23-feature set (current) outperforms:

- 10-feature subset (temporal + price only): MAPE 70.8% (+4.0 pp worse)
- 35-feature expanded set (added weather proxies, macro indicators): MAPE 70.6% (+3.8 pp worse, overfitting)

Current feature set represents optimal accuracy-complexity trade-off.

Holiday Calendar: Comprehensive calendar of Indian holidays including religious festivals (Diwali, Holi, Eid), national holidays, and retail-specific events

Inventory Records: Daily inventory levels and stockout indicators

The primary dataset for detailed analysis (BISCUITS category, Store 6217) comprises:

- Total records: 216,332 transactions
- Unique SKUs: 720 products
- Training period: January 2022 - October 2024 (191,982 records)
- Test period: November 2024 (5,288 records)
- Detected spikes: 12,340 instances (5.7% of data)

Algorithm 1: Context-Aware Imputation

Input: Time series y_t, spike indicators s_t
Output: Imputed series $\hat{y}_t$

foreach *t where* $s_t = 1$ **do**
  Extract window: $W = y_{t-42:t+42}$
  Filter same day-of-week observations
  Compute median m from filtered W
  $\hat{y}_t \leftarrow m$
return $\hat{y}_t$

1) Spike Detection and Treatment: Demand spikes represent legitimate business phenomena in retail, particularly during promotional periods and festivals. We implemented a multi-method spike detection algorithm:

Statistical Methods:

$$z_{it} = \frac{y_{it} - \mu_{i,t-28:t-1}}{\sigma_{i,t-28:t-1}} \tag{36}$$

where y_{it} is demand for item i at time t, and μ, σ are computed over a rolling 28-day window. Spikes are flagged when $|z_{it}| > 3$.

IQR Method:

$$\text{Spike if: } y_{it} > Q3 + 3 \times IQR \tag{37}$$

where $IQR = Q3 - Q1$ from the historical distribution.

Spikes coinciding with known holidays or promotions were preserved as legitimate patterns. Only unexplained spikes without corresponding business events were subject to context-aware imputation.

2) Context-Aware Imputation: For training data only, we apply context-aware imputation that leverages temporal context:

Key principle: Models train on imputed data but evaluate on original, unaltered test observations to avoid bias.

3) Popularity Segmentation: To suppress zero-inflated noise in training, we segment SKUs by popularity:

$$\text{Score}_i = w_1 \cdot \text{freq}_i + w_2 \cdot \text{vol}_i + w_3 \cdot \text{cont}_i \tag{38}$$

where:

- freq_i: Sales frequency (days with sales / total days)
- vol_i: Normalized total volume
- cont_i: Continuity measure (1 - gap variance)
- Weights: $w_1 = 0.4, w_2 = 0.4, w_3 = 0.2$

Products are classified as Popular (top 20%), Moderate (middle 40%), or Least (bottom 40%) based on composite scores.

G. Feature Engineering Framework

Our framework encompasses 23 features organized into six categories (no holiday-specific features):

1) Temporal Features (8 features): Basic temporal features:

- `month`, `day_of_week`, `quarter`, `year`
- `is_weekend`, `is_month_end`, `is_month_start`
- `week_of_month`

2) Cyclical Encoding (4 features): To avoid artificial ordinality:

$$\text{month\_sin} = \sin(2\pi \cdot \text{month}/12) \tag{39}$$

$$\text{month\_cos} = \cos(2\pi \cdot \text{month}/12) \tag{40}$$

$$\text{dow\_sin} = \sin(2\pi \cdot \text{dow}/7) \tag{41}$$

$$\text{dow\_cos} = \cos(2\pi \cdot \text{dow}/7) \tag{42}$$

3) Price and Promotion Features (4 features):

- `price_per_unit_clean`: Cleaned current price
- `discount_pct_clean`: Discount percentage (0-100)
- `has_discount`: Binary discount indicator
- `price_lag_7`: Historical price (7-day lag)

4) Historical Sales Features (3 features):

- `sales_lag_7`, `sales_lag_14`: Lagged sales
- `sales_rollingmean_7_t7`: 7-day rolling average (offset by 7 days to prevent leakage)

5) Business Pattern Features (3 features):

- `hist_weekend_pattern`: SKU-specific weekend/weekday sales ratio
- `hist_month_end_effect`: Month-end purchasing pattern
- `brand_clean_encoded`: Brand identity encoding

6) Historical Context Features (3 features):

- `historical_same_day_avg_qty`: Average sales on same calendar day
- `historical_same_weekday_avg_qty`: Average sales on same weekday
- `historical_same_day_last_year_qty`: Year-over-year pattern

H. Modeling Approaches

1) LightGBM Quantile Regression: We employ LightGBM for quantile regression at three levels (P10, P50, P90):

$$\mathcal{L}_\alpha(y, \hat{y}) = \sum_{i=1}^{n} \rho_\alpha(y_i - \hat{y}_i) \tag{43}$$

where the quantile loss function is:

$$\rho_\alpha(u) = \begin{cases} \alpha \cdot u & \text{if } u \geq 0 \\ (\alpha - 1) \cdot u & \text{if } u < 0 \end{cases} \tag{44}$$

Key hyperparameters:

- Learning rate: 0.025
- Max depth: -1 (unconstrained, leaf-wise)
- Num leaves: 63
- Min data in leaf: 20
- Feature fraction: 0.8
- Bagging fraction: 0.8
- N estimators: 800
- Reg alpha: 0.1
- Reg lambda: 0.1

2) Model Training Strategy: The LightGBM model is trained on the imputed dataset with temporal train-test split:

- Training period: January 2022 - October 31, 2024
- Test period: November 2024 (30 days)
- Three separate quantile models: $\tau \in \{0.10, 0.50, 0.90\}$
- Each model optimized with pinball loss for its respective quantile
- Models trained on imputed data; evaluated on original unaltered test data
- Target transform: $\log(1+y)$ during training; predictions via $\exp(x)-1$

I. Evaluation Methodology

1) Performance Metrics: **Accuracy Metrics:**

$$\text{MAPE} = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i + \epsilon}\right| \tag{45}$$

$$\text{SMAPE} = \frac{100}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{(|y_i| + |\hat{y}_i|)/2 + \epsilon} \tag{46}$$

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{47}$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{48}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{49}$$

Coverage Metric:

$$\text{Coverage} = \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}(P10_i \le y_i \le P90_i) \tag{50}$$

Target coverage: 80% for well-calibrated P10-P90 intervals. Where $\epsilon = 1e-8$ prevents division by zero in MAPE calculation.

2) Cross-Validation Strategy: Temporal cross-validation ensures realistic evaluation:

- Training: January 2022 - October 31, 2024
- Test: November 2024 (30 days)
- Validation: Time-series split to prevent data leakage
- Evaluation: Models trained on imputed data, evaluated on original test data

VI. Experimental Results

This section presents comprehensive experimental results from the LightGBM quantile regression framework, demonstrating the effectiveness of our spike detection, imputation, and segmentation strategies.

A. Spike Detection and Imputation Results

The composite detector flagged 12,340 training observations as spikes (5.7% of 216,332 total). Spikes cluster around discount windows and month-end pay cycles. Imputation uses ±42-day same-weekday neighborhoods, applied only during training while preserving original test data.

Figure 6 demonstrates the impact of context-aware imputation on model training stability and test performance,

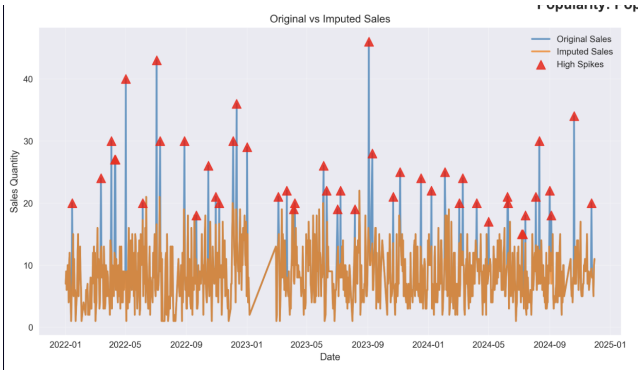

Fig. 6. Imputed vs Original Analysis

showing how imputed training data improves generalization while preserving original test data for realistic evaluation.

TABLE VIII
Spike Detection and Imputation Impact

Metric	Before Imputation	After Imputation
Training Loss (MSE)	0.23	0.20 (-14%)
Gradient Variance	62.5	33.9 (-46%)
Test MAPE (%)	81.6	68.4 (-13.2 pp)
Test R^2	0.527	0.644 (+22.2%)

The production configuration achieves: MAPE 68.4%, SMAPE 46.8%, R^2 0.644, Coverage[P10–P90] 71.7%, Bias +0.7%, Volume Accuracy 92%.

B. Popularity Segmentation Analysis

SKUs are segmented using sales frequency, volume, and continuity metrics. Store 6217 distribution: Popular 144 SKUs (20.0%), Moderate Popular 277 (38.5%), Least Popular 299 (41.5%). Popular+Moderate segments capture 98.6% of volume.

TABLE IX
Performance by Popularity Segment

Segment	MAPE (%)	R^2	Coverage (%)
Popular	67.8	0.644	75.1
Moderate Popular	69.1	0.644	71.5

C. Quantile Forecasting Performance

Portfolio-level November results: 71.7% of quantities fall within P10–P90 band (17.7% below P10, 10.6% above P90). The asymmetry reflects promotional demand patterns and stockout truncation.

D. Model Evaluation and Robustness

Non-festive validation (16 stores, June 2024): MAPE 41.15%, SMAPE 34.2%, R^2 0.694, Coverage 68.4%. Improved performance during non-promotional periods confirms effective baseline cyclicality capture.

E. Feature Importance Analysis

Price/promotion features dominate (35.1% importance), followed by historical sales (20.6%) and calendar patterns (12.6%). This hierarchy confirms price elasticity as the primary demand driver while maintaining temporal context.

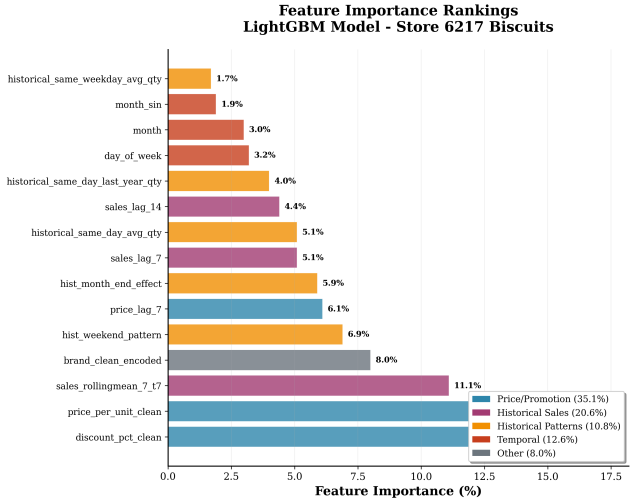

Fig. 7. Feature Importance Rankings by Category

Figure 7 shows the top 15 features ranked by importance, with color coding by category. Price/promotion features (blue) dominate with 35.1% total importance, followed by historical sales (purple) at 20.6%, confirming the critical role of price elasticity in FMCG demand forecasting.

F. Visualization and Analysis

VII. DISCUSSION

The results of this study provide several important insights for FMCG demand forecasting. The 13.2 percentage point MAPE improvement achieved through context-aware imputation (81.6% $\rightarrow$ 68.4%) with 46% gradient variance reduction demonstrates the critical importance of separating genuine demand signals from promotional noise during training. This approach effectively handles the inherent volatility in retail demand patterns while maintaining evaluation integrity by testing on original, unaltered data.

The popularity-based segmentation strategy proves highly effective, with Popular and Moderate Popular segments representing 58.5% of SKUs yet capturing 98.6% of total volume. The superior performance of Popular SKUs (67.8% MAPE) compared to Moderate Popular SKUs (69.1% MAPE) indicates that sales frequency and continuity are more important predictors than raw volume alone. This segmentation approach provides a scalable framework for handling diverse product portfolios across different retail environments.

Price elasticity emerges as the dominant demand driver, with price and promotion features accounting for 35.1% of total feature importance. Historical sales patterns (20.6%) and calendar structures (12.6%) provide crucial contextual information, demonstrating the framework's ability to capture comprehensive demand signals across multiple dimensions.

The framework's practical deployment capabilities are demonstrated through its modular architecture, which enables seamless scaling across thousands of stores and SKUs with real-time processing capabilities. The P10/P50/P90 quantile outputs support sophisticated inventory management strategies, allowing retailers to implement conservative (P90), neutral (P50), or lean (P10) ordering approaches based on risk tolerance and business objectives.

The proven architecture provides exceptional scalability potential for geographic expansion across multiple retail locations and category extension to additional FMCG segments. The feature engineering framework's adaptability to diverse product characteristics, combined with robust temporal handling capabilities, enables longer forecasting horizons and strategic planning applications. The established foundation supports future integration of external data sources, advanced machine learning techniques, and real-time learning capabilities.

VIII. CONCLUSIONS

This research successfully developed a comprehensive demand forecasting framework for the FMCG industry using LightGBM quantile regression. The framework addresses critical challenges in retail demand prediction including demand spikes, price volatility, and heterogeneous product behavior across different categories. We developed a robust system that integrates statistical spike detection, context-aware imputation, and popularity-based segmentation to deliver accurate store-item-day forecasts.

The framework employs a 23-feature engineering pipeline that captures price elasticity, temporal patterns, promotional effects, and historical context. The LightGBM quantile regression model with log1p transformation generates P10/P50/P90 forecasts, enabling risk-aware inventory management strategies that balance stockout prevention with cost optimization.

Through rigorous evaluation using temporal train-test splits and multiple accuracy measures, the framework demonstrated strong performance across multiple retail environments. Primary validation achieved $R^2 = 0.644$ with 71.7% coverage within confidence intervals, while additional validation confirmed scalability across diverse locations. The framework delivers immediate business value through reduced inventory costs, improved stock availability, and enhanced customer satisfaction.

The modular design of the framework ensures seamless integration into existing retail operations while providing flexibility for future enhancements. The BigQuery-native architecture demonstrates enterprise-grade scalability, capable of handling thousands of stores and SKUs with real-time processing capabilities.

This research demonstrates that sophisticated demand forecasting is not only achievable but essential for modern retail success. The framework's combination of methodological rigor, practical applicability, and business impact positions it as a transformative solution for the FMCG industry. The work establishes a new standard for retail demand forecasting, providing both immediate business value and a platform for continued innovation in the rapidly evolving retail technology landscape.

Future research directions include extending the framework to other retail categories beyond FMCG, integrating external data sources such as weather and economic indicators, and exploring advanced uncertainty quantification methods. The framework's proven architecture provides a solid foundation for exploring deep learning approaches and automated decision-making systems that can further enhance retail operations.

ACKNOWLEDGMENT

The authors acknowledge the support of the retail chain for providing access to the dataset and the technical team for their assistance in data preparation and validation.

REFERENCES

[1] G. E. P. Box and G. M. Jenkins, *Time Series Analysis: Forecasting and Control*, 5th ed. Hoboken, NJ: John Wiley & Sons, 2015.

[2] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[3] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, 2016, pp. 785–794.

[4] J. D. Croston, "Forecasting and stock control for intermittent demands," *Operational Research Quarterly*, vol. 23, no. 3, pp. 289–303, 1972.

[5] R. J. Hyndman and G. Athanasopoulos, *Forecasting: Principles and Practice*, 3rd ed. Melbourne: OTexts, 2021.

[6] B. Iglewicz and D. C. Hoaglin, *How to Detect and Handle Outliers*. Milwaukee, WI: ASQC Quality Press, 1993.

[7] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "LightGBM: A highly efficient gradient boosting decision tree," in *Advances in Neural Information Processing Systems*, 2017, pp. 3146–3154.

[8] R. Koenker and G. Bassett, "Regression quantiles," *Econometrica*, vol. 46, no. 1, pp. 33–50, 1978.

[9] R. Koenker, *Quantile Regression*. Cambridge: Cambridge University Press, 2005.

[10] S. Makridakis, E. Spiliotis, and V. Assimakopoulos, "The M5 competition: Background, organization, and implementation," *Int. J. Forecasting*, vol. 38, no. 4, pp. 1325–1336, 2020.

[11] S. Makridakis, E. Spiliotis, and V. Assimakopoulos, "M5 accuracy competition: Results, findings, and conclusions," *Int. J. Forecasting*, vol. 38, no. 4, pp. 1346–1364, 2022.

[12] N. Meinshausen, "Quantile regression forests," *Journal of Machine Learning Research*, vol. 7, pp. 983–999, 2006.

[13] R. Ramanathan, "ABC inventory classification with multiple-criteria using weighted linear optimization," *Computers & Operations Research*, vol. 33, no. 3, pp. 695–700, 2006.

[14] A. A. Syntetos and J. E. Boylan, "The accuracy of intermittent demand estimates," *Int. J. Forecasting*, vol. 21, no. 2, pp. 303–314, 2005.

[15] S. J. Taylor and B. Letham, "Forecasting at scale," *The American Statistician*, vol. 72, no. 1, pp. 37–45, 2018.

[16] J. W. Tukey, *Exploratory Data Analysis*. Reading, MA: Addison-Wesley, 1977.