Uncertainty-Weighted Ensembles: A Conformal Prediction Approach to Retail Sales Forecasting

Runze ZHANG   Jinghan JI   Mingyang ZHAO

Outline

Introduction

Motivation

Why Sales Forecasting Matters

- Critical for retail inventory optimization and loss minimization
- Impacts supply chain efficiency and profitability
- Challenging due to complex temporal patterns, promotions, and special events

The M5 Competition Challenge

- Predict 28 days of daily sales for Walmart items
- 30,490 hierarchical time series
- Data spans 1,913 days (Jan 2011 - Jun 2016)
- Multiple product categories, stores, and states

Our Contribution

Key Innovation: Two-Stage Framework

1. **Stage 1**: Train 113 hierarchical LightGBM models
   - Capture temporal patterns, price effects, promotions
   - Organized across state, store, category, and department levels

2. **Stage 2**: Conformal Prediction for uncertainty quantification
   - Generate calibrated prediction intervals
   - Dynamic, uncertainty-aware weights for ensemble
   - Models with tighter intervals get higher weights

Novel Approach
Integrates uncertainty quantification *directly* into ensemble aggregation

Related Work

LightGBM: Efficient Gradient Boosting

Why LightGBM?

- High performance on large-scale tabular data
- Computational efficiency for 30,490 time series
- Superior predictive accuracy

Key Techniques

1. **Gradient-based One-Side Sampling (GOSS)**:

$$\tilde{g}_i = \begin{cases} g_i & \text{if } |g_i| \geq \theta \\ \frac{1-a}{b} g_i & \text{otherwise} \end{cases}$$

2. **Exclusive Feature Bundling (EFB)**: Bundles mutually exclusive features: $O(|F| \times n) \rightarrow O(K \times n)$

LightGBM Objective Function

Training Objective at Iteration t

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \Omega(f_t)$$

where:

- $l(\cdot)$: Tweedie loss function

$$D(y, \hat{y}) = 2\left[\frac{y^{2-\rho}}{(1-\rho)(2-\rho)} - \frac{y \cdot \hat{y}^{1-\rho}}{1-\rho} + \frac{\hat{y}^{2-\rho}}{2-\rho}\right]$$

- $\hat{y}_i^{(t-1)}$: Prediction from previous iteration
- f_t: New tree being added
- $\Omega(f_t) = \gamma T + \frac{1}{2}\lambda\|w\|^2$: Regularization term

Conformal Prediction

Why Conformal Prediction?

- **Distribution-free**: No Gaussian assumptions
- **Model-agnostic**: Works with any forecasting model
- **Finite-sample guarantees**: Valid for any sample size
- Quantifies uncertainty with calibrated prediction intervals

Coverage Guarantee

Under exchangeability assumption:

$$\mathbb{P}(Y_{\text{test}} \in C(X_{\text{test}})) \geq 1 - \alpha$$

where $C(X_{\text{test}}) = [f(X_{\text{test}}) - \hat{q}, f(X_{\text{test}}) + \hat{q}]$

Conformal Prediction Procedure

1. **Split data**: Training set + Calibration set
2. **Train model**: f on training set
3. **Define nonconformity score**:

$$s(x, y) = |y - f(x)|$$

4. **Compute calibration scores**:

$$S = \{s_j = |Y_j - f(X_j)|\}_{j=1}^{n}$$

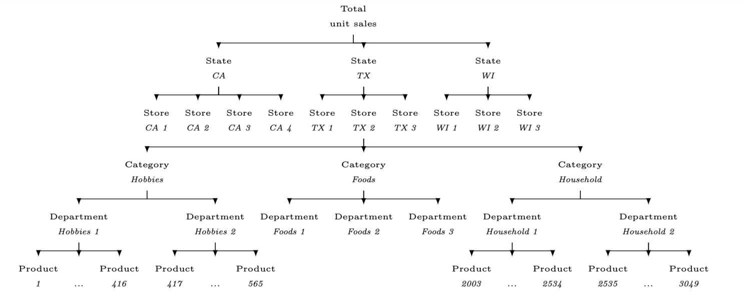
5. **Calculate quantile**:

$$\hat{q} = \text{Quantile}\left(S; \frac{\lceil (n+1)(1-\alpha) \rceil}{n}\right)$$

6. **Form prediction interval**: $C(X_{\text{test}}) = [f(X_{\text{test}}) \pm \hat{q}]$

Data Processing

Data Structure



- Sales and price data across all products
    - 3 states: California, Texas, Wisconsin
    - 10 stores across the states
    - 3 product categories: Foods, Hobbies, Household
    - 7 departments within categories
- Calendar dataset identifies special events

Exploratory Data Analysis

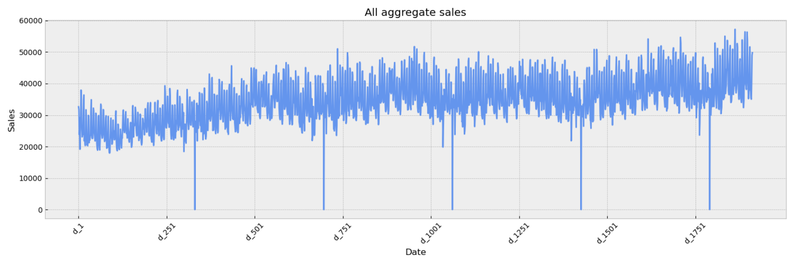
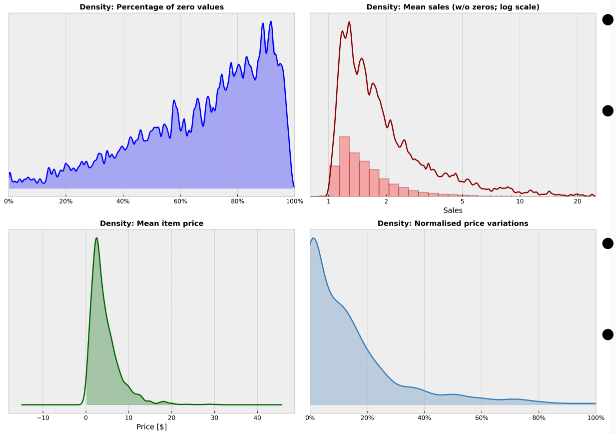


Figure 1: Aggregate sales across time

- Clear upward trend over time
- Strong weekly patterns
- Possible shorter-period overlaying seasonality

Data Characteristics



- **Extreme sparsity**: 80-90% zero-sales days
- **Heavy-tailed**: Log-normal sales distribution, concentrating around 1-2 units per day.
- **Price concentration**: Sharp peak around $5
- **Bimodal price variation**: Stable vs. promotional volatility

Feature Engineering Overview

Four Feature Categories

1. **Temporal Features**: Cyclical encodings

$$f_{\text{sin}} = \sin\left(\frac{2\pi f}{P}\right), \quad f_{\text{cos}} = \cos\left(\frac{2\pi f}{P}\right)$$

   Applied to: day of week, month, day, quarter, week

2. **Price Features**:
   - Normalized price: $\texttt{price\_norm} = \frac{\texttt{sell\_price}}{\texttt{price\_mean}+10^{-5}}$
   - `price_change`: percentage change compared to the previous day
   - `price_rolling_mean`: mean price for the past 28 days.
   - `price_momentum`: difference between current price with rolling mean price

Feature Engineering (Continued)

Four Feature Categories

3. **Lag & Rolling Window Features**:
   - Weekly lags: $\{y_{i,s,t-\ell}\}_{\ell \in \{7,14,21,28\}}$: sales of item i in store s ℓ days ago,
   - Rolling statistics: $\mu_w(t) = \frac{1}{w} \sum_{k=1}^{w} y_{i,s,t-k}$ for $w \in \{7, 14, 28\}$

4. **Meta-Model Features**:
   - Global market predictions: max, mean, std of aggregate sales
   - Inject market-wide context into local models

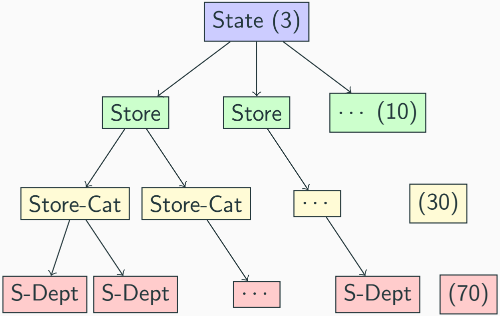
Methodology

Model Architecture Overview

Three-Component Framework

1. **Meta-Models**: Three meta-models to predict aggregate market statistics
   - $y_t^{\max} = \max_i y_{i,t}$, $y_t^{\text{mean}} = \frac{1}{N}\sum_{i=1}^{N} y_{i,t}$, $y_t^{\text{std}} = \text{std}_i(y_{i,t})$
   - Inject global context into local models
2. **Hierarchical Models (113 total)**:
   - 3 state-level models
   - 10 store-level models
   - 30 store-category models
   - 70 store-department models
3. **Uncertainty-Weighted Ensemble**: Conformal prediction weights

Hierarchical Model Structure



- Each level captures different granularity of patterns
- Coarse (state) to fine (department) specialization
- All applicable models contribute to final prediction

Uncertainty-Weighted Ensemble

Key Innovation: Dynamic Weighting by Uncertainty
For each model m at horizon h:

Step 1: Compute nonconformity scores on calibration set

$$R_j^{(m,h)} = |y_j - \hat{y}_j^{(m)}|$$

Step 2: Calculate prediction interval half-width

$$\hat{q}_\alpha^{(m,h)} = \text{Quantile}\left(\{R_j^{(m,h)}\}_{j=1}^{n_{\text{cal}}}, \frac{\lceil (n_{\text{cal}} + 1)(1-\alpha) \rceil}{n_{\text{cal}}}\right)$$

Step 3: Assign inverse-uncertainty weight

$$w_m^{(h)} = \frac{1}{(\hat{q}_\alpha^{(m,h)})^2}$$

Ensemble Aggregation Formula

Weighted Average Prediction
For item i at store s, with applicable models $\mathcal{H}(i,s)$:

$$\hat{y}_{\text{ens},i,s,t+h} = \frac{\sum_{m\in\mathcal{H}(i,s)} w_m^{(h)} \hat{y}_{i,s,t+h}^{(m)}}{\sum_{m\in\mathcal{H}(i,s)} w_m^{(h)}}$$

Typically $|\mathcal{H}(i,s)| = 4$ models (one from each level)

Advantage

- Models with tighter intervals $\to$ higher weights
- Data-driven, no manual tuning required

Recursive Multi-Step Forecasting

28-Day Horizon Strategy
For each day $h \in \{1, 2, \ldots, 28\}$:

1. **Update temporal features**: Day of week, month, cyclical encodings
2. **Meta-model prediction**: Global statistics for day $T + h$
3. **Ensemble aggregation**: Apply uncertainty-weighted combination
4. **Update lag features**:
   - $\text{lag}_d(i, s, T + h + 1) = \text{lag}_{d-1}(i, s, T + h)$
   - $\text{lag}_1(i, s, T + h + 1) = \hat{y}_{\text{ens}, i, s, T+h}$
5. **Update rolling statistics**:

$$\text{roll\_mean}_w(i, s, T + h + 1) \approx \frac{(w - 1) \cdot \text{roll\_mean}_w(i, s, T + h) + \hat{y}_{\text{ens}, i, s, T+h}}{w}$$

Results

Evaluation Metric: RMSSE

Root Mean Squared Scaled Error

$$RMSSE = \sqrt{\frac{\frac{1}{h}\sum_{t=n+1}^{n+h}\left(y_t - \hat{y}_t\right)^2}{\frac{1}{n-1}\sum_{t=2}^{n}\left(y_t - y_{t-1}\right)^2}}$$

where:

- y_t: Actual sales at time t
- $\hat{y}_t$: Forecasted sales at time t
- n: Training sample length
- h: Forecasting horizon (28 days)

Model Performance Comparison

Table 1: Performance on M5 validation set (30,490 time series)

Model	RMSE	MAE
Lightweight 40 models	1.3199	1.0531
Lightweight 110 models	1.3108	1.0455
Complete 113 models	**1.3083**	**1.0406**

Key Findings

- 110-model ensemble achieves lowest RMSSE (0.8731)
- Complete 113-model best on RMSE and MAE
- **Marginal improvements**: 0.88% RMSE, 1.19% MAE

Ablation Study: Weighting Schemes

Configuration	RMSSE	RMSE
113 + Conformal Exponential	**0.8762**	**1.3036**
113 + Conformal Inverse	0.8772	1.3083
113 + Conformal Softmax	0.8779	1.3102
113 + Equal Weight	0.8781	1.3105
110 + Conformal Softmax	0.8790	1.3172
40 + Conformal Softmax	0.8792	1.3154

Insights

- **Exponential weighting** performs best
- All conformal methods outperform equal weighting (0.22% improvement)
- 40-model ensemble only 0.34% worse than full ensemble

Key Contributions

Novel Two-Stage Framework

1. **Hierarchical ensemble**: 113 specialized LightGBM models
   - State, store, category, and department levels
   - Captures patterns at multiple granularities

2. **Conformal prediction integration**:
   - Distribution-free uncertainty quantification
   - Direct integration into ensemble aggregation
   - Dynamic, data-driven weighting