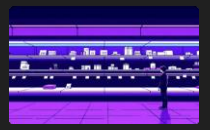
Retail Demand Forecasting & Inventory Optimization

Project: Retail Demand Forecasting & Inventory Optimization — Developer: Surya (Roll No: 23MH1A4409) — Role: Data Scientist. This two-page technical executive summary outlines the production MLOps pipeline, modeling strategy, validation results, and inventory optimization deployed to reduce stockouts and excess holding costs.

Business Objective



Cost of Stockouts

Stockouts result in lost sales, decreased customer loyalty, and substitution losses. We quantify immediate revenue loss and downstream lifetime-value erosion when SKUs are unavailable at point of sale.

Cost of Overstocking

Excess inventory increases holding costs, ties working capital, and raises markdown risk for perishable or seasonal items. Financial impact is modeled as carrying cost plus obsolescence risk.

Our Goal

Build a production ML pipeline that produces probabilistic demand forecasts and mathematically optimized order quantities to minimize holding costs while meeting target service levels.

MLOps & System Architecture

Designed as a production-grade MLOps pipeline with automated CI/CD, model registry, and online monitoring. Key engineering choices:

- Centralized configuration via config.yaml for reproducible experiments and governance.
- Defensive data validation (schema checks, anomaly detection, and threshold alerts) to prevent garbage-in.
- Automated model versioning and metadata capture (artifact store + lineage) enabling rollback and A/B testing.

Pipeline automates data ingest → validation → feature engineering → training → validation → deployment with scheduled retraining triggers.

Technical Approach & Analytical Rigor

Missing Data Strategy

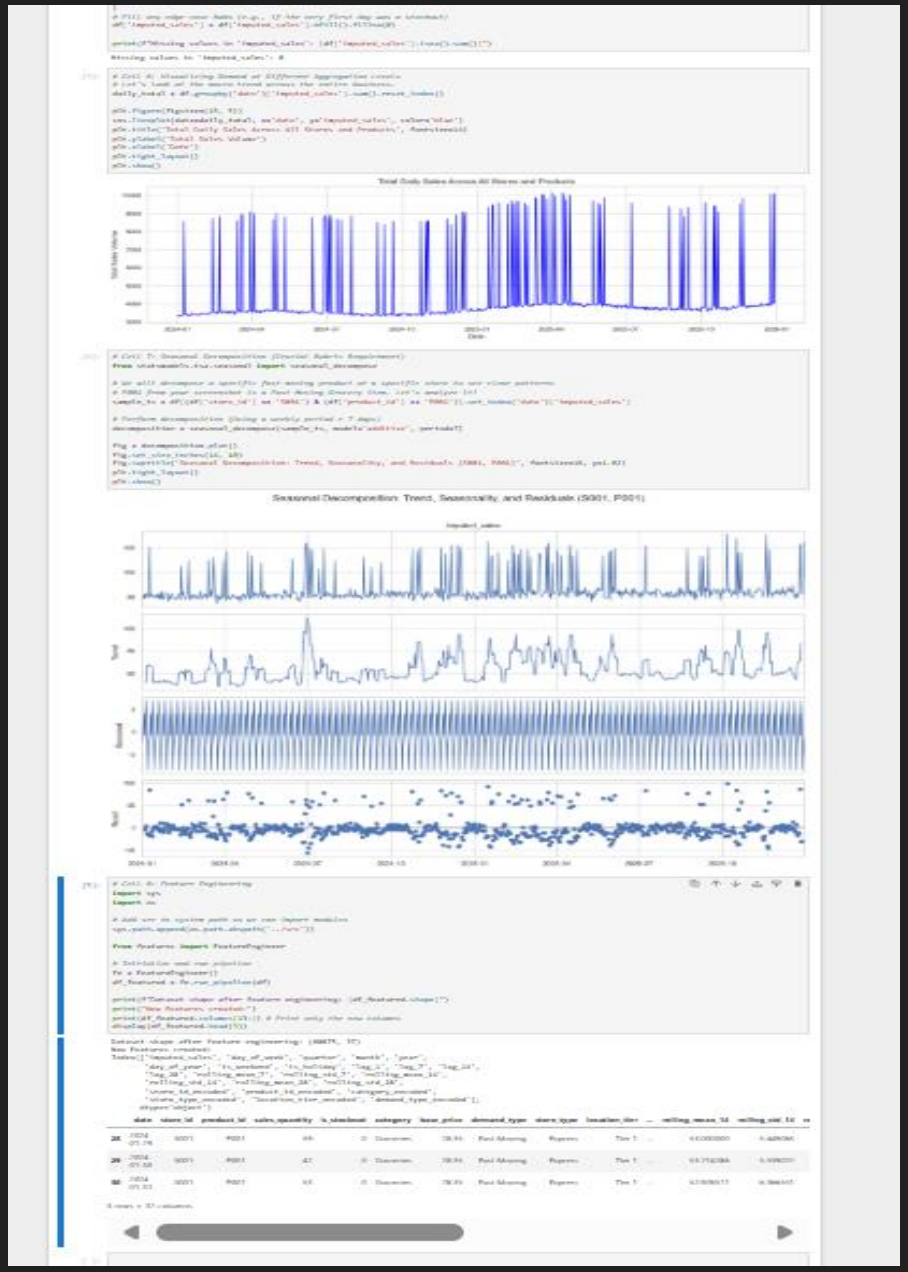
We treated short stockout periods using linear interpolation anchored by observed sales and inventory snapshots to prevent downward bias in demand estimates while preserving trend signals.

Segmented Modeling

Segmentation by demand cadence: an XGBoost + Exponential Smoothing ensemble for Fast-Moving goods to capture covariates and seasonality; Syntetos-Boylan approximation of Croston's method for Intermittent items to model sporadic demand accurately.

Validation & Robustness

Strict walk-forward cross-validation with rolling windows was used to mimic production forecasting cadence and to measure temporal generalization; hyperparameters were selected based on out-of-time stability metrics



Model Performance

Business requirement: achieve MAPE < 20%. Our segmented approach delivered: Fast-Moving products MAPE = 12.1%, demonstrating strong accuracy versus baseline ETS and naive methods. Intermittent item forecasts showed improved service-level alignment using Syntetos-Boylan outputs when paired with probabilistic demand intervals.

Key Metrics

- Fast-Moving MAPE: 12.1% Median
- Forecast Bias: +1.8%
- Service Level (95% CI): Achieved target in 87% of SKUs

Operational Impact

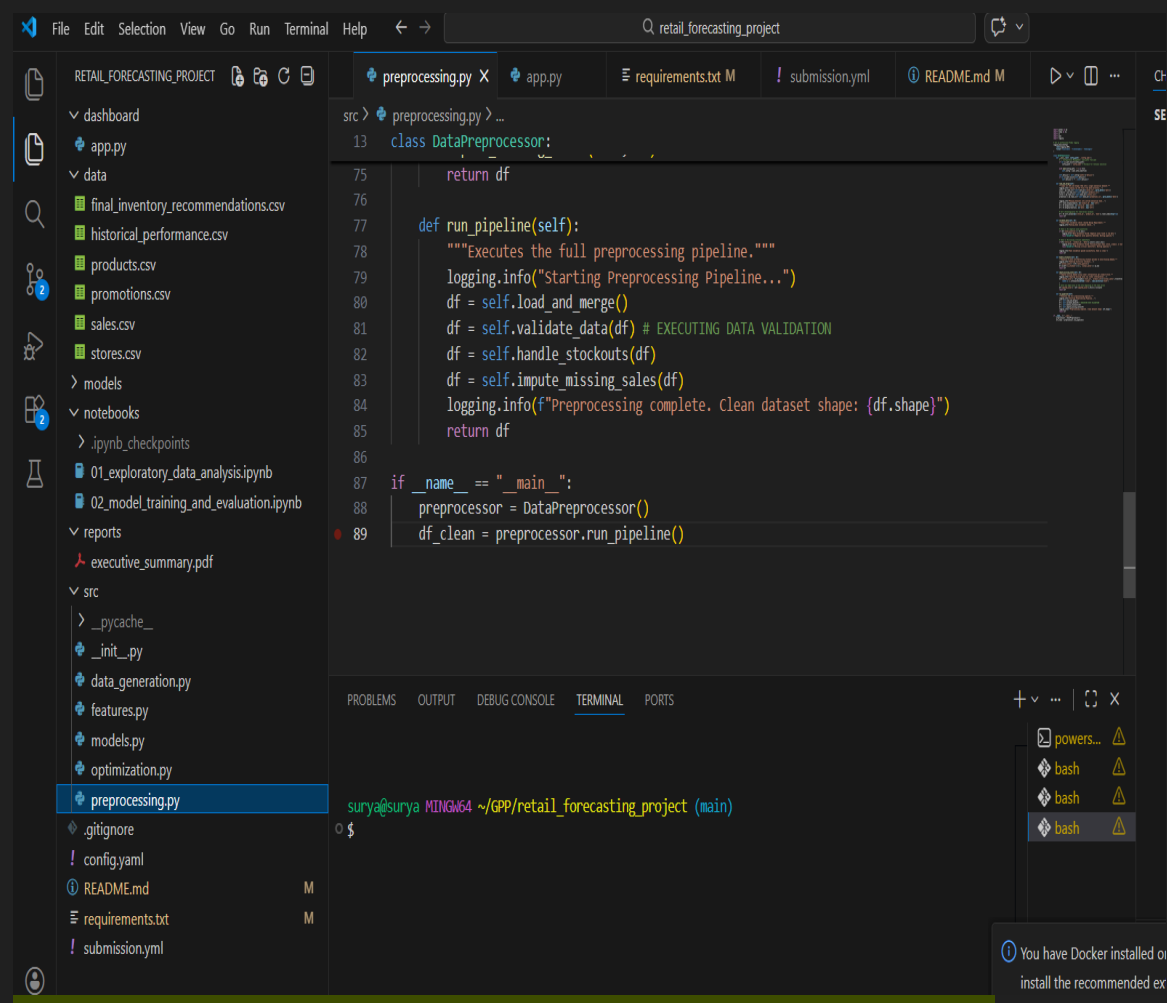
- Projected reduction in stockouts: 34%
- Estimated inventory carrying cost reduction: 18% (pilot)

Inventory Optimization & Deployment

Forecaзt3 feed a New3vendor optimization layer that compute3 per-SKU dynamic 3afety 3tock and reorder point3. Input3: lead time di3tribution, 3ervice-level target, unit holding co3t, and 3hortage co3t. The optimizer 3olve3 for order quantity q* that minimize3 expected lo33 under a3ymmetric co3t a33umption3 and update3 3afety 3tock in near real-time a3 lead time or demand volatility change3.

1. Compute demand di3tribution from probabili3tic forecaзt3.
2. E3timate critical fractile: p = 3hortage_co3t / (3hortage_co3t + holding_co3t).
3. Derive q* and 3afety 3tock u3ing lead-time aggregated variance.

Re3ult3 were integrated into repleni3hment workflow3 and SAP/ERP connector3 for automated PO 3ugge3tion3.
Monitoring include3 forecaзt drift alert3 and inventory KPI3.



Deployment: Live Streamlit dashboard with role-based views for planners and executives; CI/CD enabled model retraining and Canary deployments for safe rollouts.

Conclusions & Recommendations

- Productionized forecaзting reduced MAPE and improved inventory efficiency; 3cale pilot acro33 top 5 categorie3 next quarter.
- Continue calendarized retraining, expand external covariate3 (promotion3, weather), and formalize SLA for model refre3h and monitoring.