Universal Demand Planning via Time-Series Foundation Models

Author: Karthik Murali M

December 2025

Subject: Implementation of Amazon Chronos for Zero-Shot Supply Chain Forecasting
Repository: GitHub **Demo:** Hugging Face

1 Abstract

Traditional supply chain demand planning relies on specialized autoregressive models (e.g., XGBoost, Prophet) that require extensive feature engineering and large amounts of historical data. This technical note explores the implementation of **Amazon Chronos**, a Time-Series Foundation Model (TSFM) based on the Transformer architecture. By treating time-series data as a language modeling problem, Chronos performs *zero-shot* forecasting. Our implementation demonstrates that Chronos-Bolt achieves a 1.9% reduction in forecast error (WAPE) over a tuned XGBoost baseline on the Walmart M5 dataset without requiring domain-specific training.

2 Problem Statement

2.1 The Cold Start Challenge

Supply chains frequently encounter the *cold start* problem: new product launches (e.g., new SKUs) lack sufficient historical sales data. Conventional models such as XGBoost or ARIMA require lag-based features and therefore fail to produce reliable forecasts during early lifecycle stages, often resulting in stockouts or over-inventory during critical launch windows.

2.2 Feature Engineering Bottlenecks

Traditional forecasting pipelines spend a significant fraction of engineering effort on manual feature extraction, including rolling statistics, calendar effects, and Fourier terms. This creates technical debt and limits portability across domains such as retail, logistics, and manufacturing.

2.3 Limitations of Point Forecasts

Most production systems generate point forecasts. In supply chain management, single-value predictions are insufficient to account for volatility and uncertainty. Practitioners instead require **probabilistic forecasts** to estimate safety stock levels (e.g., the 90th percentile) to meet service-level objectives.

3 Methodology

3.1 Time-Series as a Language

Chronos reformulates time-series forecasting as a sequence modeling task using a Text-to-Text Transfer Transformer (T5) architecture. The methodology consists of three phases:

1. **Scaling:** The local context window is normalized via mean scaling to ensure scale invariance.
2. **Quantization:** Continuous values are discretized into token bins, transforming forecasting into a classification task over token probabilities.
3. **Inference:** The model autoregressively generates future tokens based on patterns learned from large-scale pretraining on diverse time-series datasets.

3.2 Implementation Pipeline

We implemented a *Universal Forecaster* capable of ingesting multiple schema formats:

- **Wide Format:** Common in retail datasets such as Walmart M5, where time steps are encoded as columns.
- **Long Format:** Common in IoT and logistics datasets, where observations are indexed by timestamps.

The inference pipeline was containerized using **Docker** to ensure environment parity and deployed via a Streamlit-based interface on Hugging Face Spaces.

4 Evaluation

4.1 Benchmark 1: Retail Demand Forecasting (Walmart M5)

Chronos-Bolt (zero-shot) was evaluated against a tuned XGBoost regressor trained with 12 lag-based features and seasonal encodings.

Model	WAPE	Training Time	Feature Engineering
XGBoost (Baseline)	0.7124	15 minutes	Manual / Heavy
Chronos-Bolt (Zero-Shot)	**0.6989**	**None**	**None**

Table 1: Retail demand forecasting results on Walmart M5

Chronos achieved a 1.9% improvement in WAPE while eliminating training and feature engineering overhead. Notably, it produced meaningful P90 safety stock estimates for intermittent-demand SKUs, where XGBoost frequently returned fractional point predictions.

4.2 Benchmark 2: Logistics Seasonality (UCI Traffic Dataset)

To evaluate generalization, Chronos was applied to hourly traffic volume data. The model successfully identified double-peak daily seasonality (morning and evening rush hours) without explicit temporal annotations, demonstrating its ability to infer complex intra-day patterns purely from tokenized sequences.

5 Limitations and Future Work

5.1 Univariate Constraints

Chronos-Bolt operates in a univariate setting and does not explicitly ingest exogenous covariates such as price or weather. Future versions (e.g., Chronos-v2) may support multivariate conditioning to further improve accuracy.

5.2 Context Window Limitations

Model performance depends on the context window length (typically 512–1024 tokens). For high-frequency data, long-term seasonal patterns may be truncated unless down-sampling strategies are employed.

5.3 Interpretability

As a Transformer-based model, Chronos lacks direct interpretability in the form of coefficients. To address this, we provide visualization of prediction intervals using Plotly to communicate uncertainty effectively to stakeholders.

6 Conclusion

This work demonstrates the practical viability of time-series foundation models for supply chain demand planning. By shifting from dataset-specific models to a zero-shot foundation model approach, organizations can reduce engineering overhead while maintaining or improving forecast accuracy. The system is particularly well-suited for environments with frequent product launches and limited historical data.