ZODIAC—ZERO-INFLATED OVERSHOOT CONTROLLED DUAL-HEAD INTEGRATION FOR ASYMMETRIC CROSS-DOMAIN FORECASTING

Igor Yakushin, Sai Krishna Kiran Beathanabhotla, Dhruv Garg & Mahmudur Rahman
Amazon.com, Inc.
{yakigor,saikrbea,gargdhru,mahmudxm}@amazon.com

ABSTRACT

Foundation models promise zero-shot forecasting across domains, yet their effectiveness for cold-start scenarios with zero-inflated distributions remains underexplored. We study cross-domain demand forecasting, predicting outcomes for items launching in new domains without historical data where a substantial fraction of launches ($\approx 30\%$) yield zero outcomes and overestimation carries asymmetric costs. We propose a specialized architecture—**ZODIAC**—combining: (1) dual-domain temporal integration via stacked recurrent layers processing source and target domain signals, (2) a dual-head design with classifier and regressor explicitly modeling zero-inflated distributions, and (3) an asymmetric loss function penalizing overestimation to align with domain-specific costs. We benchmark our approach against a pretrained in-context learner (TabPFN), an AutoML ensemble (AutoGluon), and a neural time-series model (Temporal Fusion Transformer) across six cross-domain forecasting tasks. Our model achieves 80% WAPE, a 13% relative improvement over TabPFN, 25% over AutoGluon, and 26% over TFT while reducing systematic overprediction from 66–87% to under 41%, a property unachievable with models lacking loss customization.

Track: Research Track

1 INTRODUCTION

Foundation models and large-scale pretrained systems, such as Chronos (Ansari et al., 2024), TimesFM (Das et al., 2024) for time series, and TabPFN (Hollmann et al., 2023) for tabular prediction have demonstrated impressive zero-shot forecasting across diverse domains. However, their effectiveness under *structural domain constraints*, such as zero-inflated distributions, asymmetric error costs, and cold-start scenarios remains underexplored. Such constraints naturally co-occur whenever forecasting must generalize from one or more established source domains to a novel target domain (e.g., store, marketplace, branch) with no prior observations. Cross-border e-commerce demand forecasting exemplifies such a setting: when sellers launch products in new marketplaces, 70–80% of candidates lack target marketplace history, 15–25% of launches yield zero sales at the 90/365-days horizon, and overestimation carries disproportionate costs (excess inventory, capital misallocation).

Existing approaches each fail on at least one axis. Pretrained in-context learners such as TabPFN and AutoML ensembles such as AutoGluon-Tabular (Erickson et al., 2020) offer strong general-purpose prediction but lack loss customization for asymmetric costs or zero-inflated targets. Neural time-series models such as TFT (Lim et al., 2021) handle temporal structure but assume well-behaved continuous distributions. None jointly address the cold-start, zero-inflation, and asymmetric cost challenges that co-occur in cross-domain forecasting.

We propose **ZODIAC**, a lightweight architecture with three design choices targeting the challenges above: (1) **dual-domain temporal integration**—stacked LSTMs process time series from both source and target domains, providing proxy demand signals for cold-start items lacking target-marketplace history (we favour LSTMs over transformers as our sequences are short, $T{=}12$ monthly steps, and the model must score 42M+ items weekly; see Section 2); (2) **dual-head architecture**—a classifier predicts non-zero demand probability while a regressor estimates conditional magnitude,

explicitly decomposing the zero-inflated distribution; and (3) **asymmetric loss**—a composite BCE + asymmetric MSE loss with tunable penalty γ enables direct control over the overprediction rate, a capability absent in foundation models with fixed symmetric losses.

Our key contributions are: (i) a systematic benchmark of pretrained, AutoML, and neural time-series models on cross-domain cold-start forecasting with zero-inflated distributions across six real-world marketplace arcs from a major global e-commerce platform; (ii) a specialized architecture whose domain-specific design choices yield 13–26% relative WAPE improvement over all baselines while reducing overprediction from 66–87% to 38–41%; and (iii) empirical evidence that general-purpose models cannot control prediction bias or handle zero-inflated targets, identifying a regime where lightweight specialized architectures remain necessary complements to foundation models.

2 METHOD

We present a general neural framework for cold-start demand forecasting under zero-inflated distributions with asymmetric error costs. The framework is applicable to any setting where: (a) items must be scored in a target domain without historical observations, (b) proxy signals can be constructed from one or more source domains and from similar items in the target domain, and (c) a significant fraction of outcomes are zero. We instantiate this framework for cross-border e-commerce, where sellers launch products into new international marketplaces; however, the same source-to-target structure applies to new store openings, and new branch forecasting.

Problem Formulation. Consider a source domain $\mathcal{S}$ and target domain $\mathcal{T}$. For each candidate item o_i observed in $\mathcal{S}$, we predict a non-negative outcome $\hat{y}_i \geq 0$ in $\mathcal{T}$ over a specified horizon h. The prediction function takes three input modalities: $\hat{y}_i = f_\theta\Big(\mathbf{x}_i^{(s)},\ \mathbf{x}_i^{(t)},\ \mathbf{x}_i^{(\text{static})}\Big)$, where $\mathbf{x}_i^{(s)} \in \mathbb{R}^{K\times T}$ is a multivariate time series of K metrics over T time steps from the source domain, capturing the item's historical trajectory; $\mathbf{x}_i^{(t)} \in \mathbb{R}^{K\times T}$ is an aggregated time series from *similar items* in the target domain, serving as a proxy for unobserved target-domain history (the cold-start signal); and $\mathbf{x}_i^{(\text{static})} \in \mathbb{R}^d$ encodes static attributes of the item and its context (e.g., item properties, seller characteristics). The key modeling challenge is that the target variable y_i follows a zero-inflated distribution: $P(y_i = 0) = \pi > 0$, with the positive component $y_i \mid y_i > 0$ being continuous and right-skewed. Standard regression losses treat zero and non-zero outcomes identically, leading to poor calibration. Furthermore, in many applications the cost of overestimation exceeds that of underestimation (e.g., excess inventory vs. missed opportunity), requiring asymmetric loss design.

Target-domain proxy construction. Since o_i has no history in $\mathcal{T}$, we construct $\mathbf{x}_i^{(t)}$ by identifying similar items in $\mathcal{T}$ and aggregating their temporal signals. We employ a hierarchical similarity strategy with progressively relaxed matching criteria: (1) embedding-based semantic similarity with score-weighted aggregation, (2) category and price-bin matching with uniform aggregation, and (3) category-only matching. This cascade maximizes coverage—the fraction of items for which proxy signals can be constructed—while prioritizing signal quality when tighter matches are available.

Architecture. The architecture (Figure 1) consists of four components designed to handle the heterogeneous input modalities and zero-inflated output distribution. ***Temporal encoder.*** The concatenated source and target time series $[\mathbf{x}_i^{(s)};\ \mathbf{x}_i^{(t)}] \in \mathbb{R}^{2K\times T}$ are processed by a stacked LSTM (Hochreiter & Schmidhuber, 1997) with L layers and hidden dimension H, producing a fixed-length temporal representation $\mathbf{h}_i \in \mathbb{R}^H$ from the final hidden state. The LSTM captures sequential dependencies and seasonality patterns across both domains jointly, allowing the model to learn cross-domain temporal correlations. ***Feature fusion.*** The temporal representation is concatenated with static features and processed through a sequence of fully-connected layers with ReLU activations and dropout regularization: $\mathbf{z}_i = \text{FC}_M \circ \cdots \circ \text{FC}_1\Big([\mathbf{h}_i;\ \mathbf{x}_i^{(\text{static})}]\Big) \in \mathbb{R}^D$ producing a shared representation $\mathbf{z}_i$ of dimension D that captures interactions between temporal dynamics and static attributes. ***Dual-head decoder.*** The shared representation feeds two parallel output heads that decompose the zero-inflated prediction: *Classifier head*: A linear layer with sigmoid activation outputs $p_i = P(y_i > 0 \mid \mathbf{z}_i)$, the probability of non-zero outcome. *Regressor head*: A linear layer with ReLU activation outputs $\hat{y}_i = \mathbb{E}[y_i \mid y_i > 0, \mathbf{z}_i]$, the conditional magnitude given non-zero outcome. The final point prediction is $p_i \cdot \hat{y}_i$, which naturally produces conservative estimates: even when the regressor predicts high magnitude, a low classifier probability attenuates the forecast. This decomposition is motivated

by classical two-part models (Belotti et al., 2015) and their neural extensions (Kong et al., 2020), adapted here for the cold-start setting with dual-domain temporal inputs.

Composite Asymmetric Loss. We design a loss function that jointly trains both heads while encoding the asymmetric cost structure: $L = \alpha\, L_{\text{BCE}} + (1-\alpha)\, L_{\text{aMSE}}$.

The **classification component** trains the occurrence head via binary cross-entropy:

$$L_{\text{BCE}} = -\frac{1}{N}\sum_{i=1}^{N}\left[z_i \log p_i + (1-z_i)\log(1-p_i)\right], \quad z_i = \mathbb{I}[y_i > 0] \tag{1}$$

The **regression component** trains the magnitude head on non-zero instances $\mathcal{N} = \{i : y_i > 0\}$ with an asymmetric penalty:

$$L_{\text{aMSE}} = \frac{1}{\sigma^2 + \epsilon} \cdot \frac{1}{|\mathcal{N}|}\sum_{i\in\mathcal{N}} \ell_i, \quad \ell_i = \begin{cases} \gamma(\hat{y}_i - y_i)^2 & \text{if } \hat{y}_i > y_i \\ (\hat{y}_i - y_i)^2 & \text{if } \hat{y}_i \le y_i \end{cases} \tag{2}$$

Three hyperparameters control the loss behavior. The *asymmetry parameter* $\gamma > 1$ penalizes overestimation relative to underestimation; by tuning γ, practitioners can directly control the overshoot rate (fraction of predictions exceeding actuals) to match domain-specific cost asymmetries. The *scale normalization* $\sigma^2 = \text{Var}(y_i \mid y_i > 0)$ ensures the regression loss adapts to the magnitude of the target variable, enabling stable training across domains with different demand scales. The *mixture weight* $\alpha \in [0, 1]$ balances the relative importance of correctly identifying zero outcomes versus accurately predicting non-zero magnitudes. This loss design addresses a fundamental limitation of foundation models: since pretrained models use fixed symmetric losses (typically MSE or likelihood-based), they cannot optimize for domain-specific cost asymmetries without retraining or fine-tuning—which may not be feasible or may degrade their general-purpose capabilities.

Uncertainty Quantification via Monte Carlo Dropout. For risk-sensitive decisions we employ Monte Carlo (MC) Dropout (Gal & Ghahramani, 2016): at inference, B stochastic forward passes with dropout enabled yield an ensemble $\{\hat{y}_i^{(b)}\}_{b=1}^{B}$ from which we extract confidence intervals (e.g., $[q_{0.25}, q_{0.75}]$) and distributional statistics. Unlike quantile regression (Koenker & Bassett, 1978) or deep ensembles, MC Dropout requires a single trained model and produces many quantiles at inference time, scaling to millions of items without additional training cost.

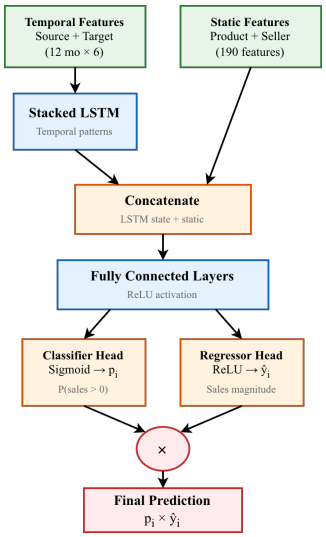

Figure 1: Model architecture. Dual-domain temporal features are encoded via stacked LSTMs, fused with static features through shared dense layers, and decoded through parallel classifier and regressor heads.

3 Experiments

Dataset. We evaluate on a large-scale proprietary dataset spanning six source-to-target arcs (US→{UK, DE, FR, IT, ES, JP}) over 2 years, with 250K+ training offers per arc, where an *offer* denotes one product launched by one seller in a marketplace. Each offer is described by K=3 monthly time series from the source marketplace (units sold, gross merchandise sales, page views) and K=3 aggregated from similar offers in the target marketplace, each over T=12 months, plus d=190 static features (product category, price, brand, seller tenure, fulfillment method, reputation metrics). The dataset poses three simultaneous challenges: cold-start (70–80% of offers lack target-marketplace history), zero-inflation (15–25% at the 90-day horizon, 5–10% at 365 days), and right-skewed demand distributions that vary across geographies.

Baselines. We compare against TabPFN (Hollmann et al., 2023) (pretrained in-context learner), AutoGluon-Tabular (Erickson et al., 2020) (AutoML ensemble), and TFT (Lim et al., 2021) (neural time-series model). TabPFN and AutoGluon treat temporal features as flat tabular inputs; TFT processes them as ordered sequences.

Table 1: 365-day forecast: WAPE (%) and Overshoot rate (%). Lower WAPE is better; overshoot near 50% is ideal. Best results in **bold**.

	TabPFN		AutoGluon		TFT		**ZODIAC (Ours)**	
Arc	WAPE	Overshoot	WAPE	Overshoot	WAPE	Overshoot	**WAPE**	**Overshoot**
US→UK	91	73	88	73	100	82	**81**	**41**
US→DE	91	72	121	83	96	81	**81**	**38**
US→JP	91	73	102	81	108	87	**80**	**39**
US→FR	94	70	111	78	124	86	**81**	**38**
US→IT	94	71	101	66	113	84	**80**	**39**
US→ES	91	69	111	83	110	84	**78**	**39**
Avg	92	71	106	77	109	84	**80**	**39**

Metrics. We report WAPE (Weighted Absolute Percentage Error, WAPE $= \frac{\sum_{i=1}^{N} |y_i - \hat{y}_i|}{\sum_{i=1}^{N} y_i} \times 100\%$; lower is better) and overshoot rate (fraction of predictions exceeding actuals, $\frac{1}{N}\sum_{i=1}^{N} \mathbb{I}[\hat{y}_i > y_i] \times 100\%$. Values near 50% indicate unbiased predictions; 50% is unbiased).

Training. Our model is trained per arc-horizon with Bayesian HPO over 200 configurations ($\sim$5 min/run, $\sim$250K parameters). TabPFN was limited to 20K-record context with an ensemble of 10 predictors. AutoGluon was given a 2-hour search budget. TFT used a single default configuration (early stopping, patience=4) as each run took $\sim$4 hours, making extensive HPO infeasible.

Results. Table 1 shows that ZODIAC achieves 78–81% WAPE across all six arcs, a 13% relative improvement over TabPFN (avg. 92%), 25% over AutoGluon (avg. 106%), and 26% over TFT (avg. 109%), with consistent gains across geographically diverse markets. All baselines exhibit systematic overprediction (overshoot 66–87%); our asymmetric loss with tunable γ reduces overshoot to 38–41%, a calibration impossible with TabPFN (no gradient-based training) and impractical with AutoGluon (no loss customization). TFT performs worst despite its temporal design (avg. WAPE 109%, overshoot 84%), likely because short 12-step sequences and zero-inflation violate its continuous-target assumption. By jointly processing source-domain and target-domain time series, the LSTM encoder captures cross-domain correlations that flat tabular representations discard.

4 Ablation Study

ZODIAC combines two design choices: a *dual-head architecture* that decomposes predictions into a zero/non-zero classifier and a magnitude regressor, and an *asymmetric loss* that penalizes overestimation more than underestimation. We ablate each component to quantify its contribution.

Dataset. Since the proprietary cross-border data (Section 3) cannot be released, we conduct ablations on a public benchmark derived from the M5 Forecasting dataset Makridakis et al. (2020) following Igor Yakushin et al. (2026). The M5 data is transformed to reproduce three structural properties of our production setting: (i) *cold-start*: labels come from a different store than the input features, simulating cross-domain prediction; (ii) *zero-inflation*: 70% of labels are synthetically zeroed, matching the zero-inflated distributions in production; and (iii) *high-dimensional static features*: one-hot encoding produces d=160 static inputs alongside T=12 monthly time steps.

Setup. We evaluate a 2×2 factorial over two factors: **architecture**—single-head (α=0, regressor only: $\hat{y}_i = r_i$) vs. dual-head (α=0.5: $\hat{y}_i = p_i \cdot r_i$); and **loss symmetry**—symmetric (γ=1, standard MSE) vs. asymmetric (γ=100, overestimation penalized 100×). All other hyperparameters are fixed. We report WAPE (%, lower is better), overshoot rate (%, fraction of predictions exceeding actuals; 50% is unbiased), and R^2.

Results. Figure 2 visualizes the WAPE–overshoot trade-off on the held-out test set. The dual-head architecture is the most impactful component: adding the classifier head under symmetric loss (ⓐ→ⓒ) reduces WAPE from 76% to 47% (38% relative, R^2: 0.833 → 0.860), as low p_i attenuates forecasts for likely-zero items. The asymmetric loss independently controls overestimation: increasing γ reduces overshoot from 80% to 65% for single-head (ⓐ→ⓑ) and from 58% to 48% for dual-head (ⓒ→ⓓ), approaching the unbiased ideal of 50%. The two components interact: asymmetric loss improves both WAPE and overshoot for the single-head model (WAPE 76 → 64), but for the dual-head model it reduces overshoot (58 → 48) while increasing WAPE (47 → 57, R^2: 0.860 → 0.760), since the classifier already handles zero-attenuation and the additional penalty over-suppresses non-zero forecasts. This complementarity motivates per-arc γ tuning in production

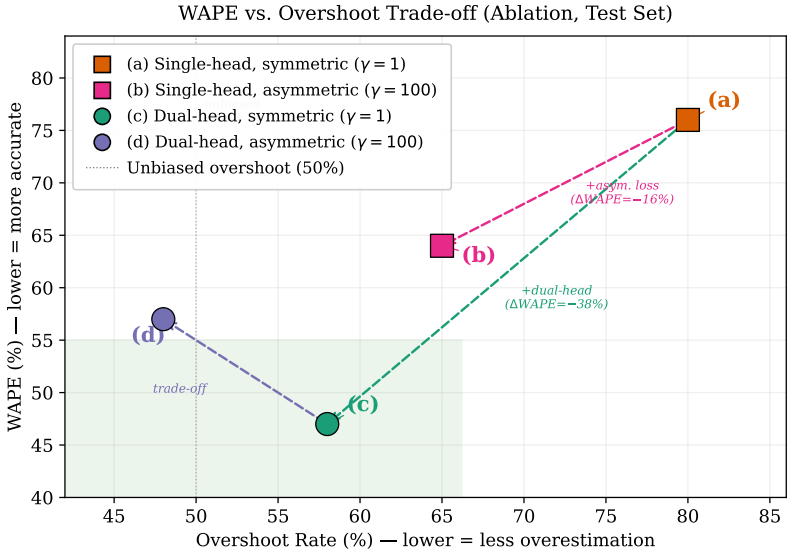

Figure 2: Ablation: WAPE vs. overshoot on the test set. Squares: single-head; circles: dual-head. Dashed arrows show component effects. Adding the dual-head (ⓐ→ⓒ) gives the largest WAPE gain; adding asymmetric loss (ⓒ→ⓓ) reduces overshoot toward 50% (dotted line) at higher WAPE.

(Section 3). Test WAPE on this benchmark (47–76%) is lower than on the proprietary data (78–81%, Table 1), reflecting stronger feature–label correlations; these ablations establish directional component contributions.

5 Related Work

Time-series foundation models (Ansari et al., 2024; Das et al., 2024; Rasul et al., 2024; Woo et al., 2024) and tabular in-context learners such as TabPFN (Hollmann et al., 2023) achieve strong zero-shot performance but assume continuous targets and symmetric losses, lacking mechanisms for zero-inflated outcomes or asymmetric costs. Architectural advances such as SPADE (Wolff & Baumann, 2024) improve forecasting via spectral attention but still require historical data. For cold-start settings, GNN-based approaches (Panagopoulos et al., 2023) transfer knowledge across similar products, RNN clustering methods (Bandara et al., 2020) share parameters across related series, and content-based strategies from recommendation systems (Schein et al., 2002; Volkovs et al., 2017) leverage item features; however, none addresses zero-inflated demand or asymmetric costs. Classical hurdle models (Belotti et al., 2015) and neural two-part extensions (Kong et al., 2020) decompose zero-inflated distributions into occurrence and magnitude components but do not incorporate cross-domain temporal signals. ZODIAC unifies these threads: it transfers source-domain temporal knowledge to cold-start targets via dual-domain integration, models zero-inflation through a dual-head architecture, and controls overprediction with an asymmetric loss, jointly addressing challenges not considered together in prior work.

6 Conclusion

ZODIAC demonstrates that lightweight, purpose-built architectures encoding domain structure—zero-inflation via dual-head decomposition, asymmetric costs via tunable γ, and cold-start via dual-domain temporal integration—can substantially outperform general-purpose foundation models on specialized forecasting tasks. Across six cross-border arcs, ZODIAC achieves 13–26% relative WAPE improvement over TabPFN, AutoGluon, and TFT while reducing overshoot from 66–87% to under 41%, with a compact footprint ($\sim$250K parameters, $\sim$5 min training) scaling to 42M+ items weekly in production. Our evaluation is limited to cross-border e-commerce and does not benchmark time-series foundation models (Chronos, TimesFM), which expect univariate sequential input incompatible with our multimodal feature space (d=190 static + $2K$=6 temporal channels). Future work will extend ZODIAC to aggregate signals from multiple source domains $[\mathcal{S}_1, \ldots, \mathcal{S}_n]$ via attention-based fusion, explore lightweight temporal attention encoders for domains with differing seasonality, and validate the source-to-target framework on additional settings such as new store openings and market entry planning. These results suggest complementary rather than substitutive roles for specialized and foundation models.

REFERENCES

Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, Jasber Zughaibi, Danielle Maddix, Michael Mahoney, Kari Torkkola, Andrew Gordon Wilson, Michael Bohlke-Schneider, and Yuyang Wang. Chronos: Learning the language of time series. *arXiv preprint arXiv:2403.07815*, 2024.

Kasun Bandara, Christoph Bergmeir, and Slawek Smyl. Forecasting across time series databases using recurrent neural networks on groups of similar series: A clustering approach. *Expert Systems with Applications*, 140:112896, 2020.

Federico Belotti, Partha Deb, Willard Manning, and Edward Norton. Twopm: Two-part models. *Stata Journal*, 15:3–20, 2015.

Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024.

Nick Erickson, Jonas Mueller, Alexander Shirkov, Hang Zhang, Pedro Larroy, Mu Li, and Alexander Smola. Autogluon-tabular: Robust and accurate automl for structured data. *arXiv preprint arXiv:2003.06505*, 2020.

Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, 2016.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8): 1735–1780, 1997.

Noah Hollmann, Samuel Müller, Katharina Eggensperger, and Frank Hutter. TabPFN: A transformer that solves small tabular classification problems in a second. In *International Conference on Learning Representations (ICLR)*, 2023.

Igor Yakushin et al. M5 Dataset Modification Scripts for ICLR. `https://www.github.com/M5_modification_for_iclr`, 2026.

Roger Koenker and Gilbert Bassett. Regression quantiles. *Econometrica*, 46(1):33–50, 1978.

Shufeng Kong, Junwen Bai, Jae Hee Lee, Di Chen, Andrew Allyn, Michelle Stuart, Malin Pinsky, Katherine Mills, and Carla P. Gomes. Deep hurdle networks for zero-inflated multi-target regression: Application to multiple species abundance estimation. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI)*, 2020.

Bryan Lim, Sercan Ö. Arık, Nicolas Loeff, and Tomas Pfister. Temporal fusion transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting*, 37(4): 1748–1764, 2021.

Spyros Makridakis, Evangelos Spiliotis, and Vassilios Assimakopoulos. M5 Forecasting - Accuracy. `https://www.kaggle.com/competitions/m5-forecasting-accuracy`, 2020. Kaggle Competition. Estimate the unit sales of Walmart retail goods.

George Panagopoulos, Fragkiskos D. Malliaros, and Michalis Vazirgiannis. Graph neural networks for demand forecasting in e-commerce. In *Proceedings of the ACM Web Conference*, 2023.

Kashif Rasul, Arjun Ashok, Andrew Robert Williams, Arian Khorasani, George Adamopoulos, Rishika Bhatt, Sun Peng, Christos Faloutsos, and Michael Bohlke-Schneider. Lag-llama: Towards foundation models for probabilistic time series forecasting. In *International Conference on Learning Representations (ICLR)*, 2024.

Andrew I. Schein, Alexandrin Popescul, Lyle H. Ungar, and David M. Pennock. Methods and metrics for cold-start recommendations. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 253–260, 2002.

Maksims Volkovs, Guangwei Yu, and Tomi Poutanen. Dropoutnet: Addressing cold start in recommender systems. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017.

Moritz Wolff and Philipp Baumann. SPADE: Spectral decomposition for time series forecasting with attention. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024.

Gerald Woo, Chenghao Liu, Akshat Kumar, Caiming Xiong, Silvio Savarese, and Doyen Sahoo. Unified training of universal time series forecasting transformers. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024.