Midterm Report: Conditional GAN-based Sales Prediction Using Reviews and Sales Data

Team: Li Siqi (122090271), Li Chudikang (122040057), Yichun Wang (122090550)

Introduction

We aim to improve short-term sales forecasting by incorporating customer reviews and temporal signals into a conditional generative framework. Traditional models trained solely on historical sales can lag during rapid shifts caused by sentiment changes, promotions, or holidays. Our project proposes a two-stage approach: first, train a conditional GAN (cGAN) to generate realistic sales sequences conditioned on review sentiment, time features and so on; second, use these synthetic sequences to augment training of forecasters. This is useful for business that need timely responsiveness to market sentiment while mitigating data sparsity and regime change issues.

Research Paper

Yoon, J., Jarrett, D., & van der Schaar, M. (2019). Time-series generative adversarial networks. In Advances in Neural Information Processing Systems (NeurIPS) 32 (pp. 5509–5519).
https://papers.neurips.cc/paper/8789-time-series-generative-adversarial-networks.pdf

Lim, B., Arik, S. Ö., Loeff, N., & Pfister, T. (2021). Temporal fusion transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting*, *37*(4), 1748–1764. https://doi.org/10.1016/j.ijforecast.2021.03.012

Salinas, D., Flunkert, V., Gasthaus, J., & Januschowski, T. (2020). DeepAR: Probabilistic forecasting with autoregressive recurrent networks. *International Journal of Forecasting*, *36*(3), 1181–1191. https://doi.org/10.1016/j.ijforecast.2019.07.001

Yu, L., Zhang, W., Wang, J., & Yu, Y. (2017). SeqGAN: Sequence generative adversarial nets with policy gradient. In *Proceedings of the 5th International Conference on Learning Representations (ICLR)*. arXiv. https://arxiv.org/abs/1609.05473

McAuley, J., & Leskovec, J. (2013). Hidden factors and hidden topics: Understanding rating dimensions with review text. In *Proceedings of the 7th ACM Conference on Recommender Systems (RecSys)* (pp. 165–172). ACM. https://doi.org/10.1145/2507157.2507162

Oreshkin, B. N., Carpov, D., Chapados, N., & Bengio, Y. (2020). N-BEATS: Neural basis expansion analysis for interpretable time series forecasting. In *Proceedings of the 8th International Conference on Learning Representations (ICLR)*. arXiv.

https://arxiv.org/abs/1905.10437

Data Collection

We choose the Olist Brazilian E-Commerce Public Dataset which contains detailed transactional data across multiple relational tables. These include information on customer orders, order items, star-rated and timestamped textual reviews and product-level metadata. To facilitate modeling at the intersection of time and product, we reorganize the raw data into a product-day panel structure. Each row in this panel represents a specific product on a given day and records the number of units sold, the number of reviews received, the average star rating and a rolling window of review texts aligned with that day's activity.

In addition to the Olist dataset, we optionally incorporate two auxiliary data sources. The first is the Yelp review dataset. Although it doesn't tied to sales figures, it offers a rich corpus of annotated review texts. The second is a public calendar of Brazilian holidays, which we use to flag national holidays and other special days. This allows us to account for irregular demand fluctuations driven by seasonal trends or promotional campaigns.

PData reprocessing

ATo prepare the dataset for time-series modeling, we first aggregate daily sales at the product level and retain only those products with sufficiently long activity histories to ensure that each sequence carries enough temporal signal for learning. Customer review events are aligned to calendar days so that their content and associated ratings can be used to construct day-level features. From these, we derive temporal covariates such as day-of-week, month, and holiday indicators, along with moving averages and lag-based statistics over windows of 7, 14, and 30 days, all of which serve as input signals for the forecasting models.

To standardize across products with varying sales scales, we experiment with z-score normalization and robust scaling methods. Missing days, where no sales or reviews occur, are handled via explicit zero imputation or forward-fill indicators depending on the modeling approach. Padding is applied where necessary and masked accordingly to prevent information leakage across samples. Within each training window, we reserve the final temporal block per product as a validation set, thereby preserving the natural chronological order of events and avoiding look-ahead bias.

Method

Conditional generator (cGAN with WGAN-GP)

Condition y = [encoded past sales window (e.g., 30 days), review sentiment embeddings and review statistics, temporal features for the next 7 days].

Sales encoder: LSTM over past 30-day sales and exogenous covariates to produce temporal context.

Review encoder: BERT-derived daily embeddings aggregated over the lookback window; include star ratings and review counts.

Time encoder: learned embeddings or one-hot for day-of-week, holiday flags, promo indicators if available.

Generator G(z|y): seq2seq network (LSTM/GRU with attention or lightweight Transformer decoder) that outputs a 7-day sales sequence.

Discriminator D(x|y): temporal CNN or bidirectional LSTM that scores realism of 7-day sequences given y.

Loss: WGAN-GP with gradient penalty; additional reconstruction/consistency losses (optional) to tether generated sequences to plausible scale ranges.

Downstream forecasting

Strategy A (augmentation): Use real + high-quality synthetic (G(z|y)) sequences to augment training of deterministic/probabilistic forecasters (LSTM/GRU, TFT, and gradient-boosted trees with lagged features).

Strategy B (direct generative forecasting): Use G(z|y) to produce multiple stochastic trajectories and summarize into quantiles/means.

Baselines

Naive/seasonal: last-week repeat, moving-average.

Classical: ARIMA/SARIMA where feasible; Prophet with holidays.

Machine learning: XGBoost/LightGBM on lag and calendar features.

Deep learning: LSTM/GRU with covariates; DeepAR; Temporal Fusion Transformer.

Evaluation

Primary forecasting metrics:

MAE, RMSE, sMAPE, MAPE (with care on zero-sales days), WAPE/ND.

Probabilistic: pinball loss at $\tau \in \{0.1, 0.5, 0.9\}$; interval coverage for 80%/90% prediction intervals.

Aggregation: per-product then macro-average; also category-level aggregation to assess business relevance.

S**econdary/business metrics:**

Event responsiveness: performance during holiday weeks or sentiment shocks (days with large negative-review spikes).

Cold-start sensitivity: shorter history products vs well-established ones.

E**xperiments**

E1: Data augmentation effectiveness

Compare each baseline trained on real-only vs real+synthetic. Measure change in MAE/sMAPE; analyze where augmentation helps (low data regime, volatile items).

E2: Value of reviews

Ablate review features: sales-only vs sales+ratings vs sales+ratings+text embeddings.

E3: Horizon generalization

Forecast horizons 1, 3, 7, 14 days. Assess how cGAN benefits vary with horizon.

E54 Cold-start/short-history

Train on subsets with limited history (e.g., $60 - 90$ days) to test augmentation benefits.

C**urrent Progress**

Data:

Olist dataset downloaded and integrated; product-day panel built.

Review text preprocessing pipeline implemented with multilingual BERT; sentiment embeddings cached per day; star rating and review count features aligned.

Holiday flags added; lag and moving-average features generated.

Baselines:

Prophet and naive baselines implemented and validated on rolling-origin splits.

LSTM baseline with covariates implemented (PyTorch Lightning); early runs confirm stable training and sensible seasonality capture.

cGAN:

First pass of conditional WGAN-GP implemented with LSTM encoders; discriminator shows training stability with gradient penalty.

Eooling:

Reproducible pipelines with Hydra/Lightning; experiment tracking in Weights & Biases;

GPU training enabled.

Preliminary Observations (qualitative)

Planned Improvements and Next Steps

Architecture refinements:

Replace sales encoder with a lightweight Transformer to better capture long-range dependencies.

Try Temporal Convolutional Networks for the discriminator to stabilize gradients.

Quantile loss fine-tuning for downstream forecasters to better align with probabilistic evaluation.

Dategory-aware generation to reduce mismatch between item types.

Risks, Failures, and Mitigations

AN instability and mode collapse: using WGAN-GP, spectral normalization, feature matching, and early stopping on validation discriminator metrics.

Sparse or noisy reviews: backoff to star-ratings and review counts; smoothing windows; leverage multilingual models suited for Portuguese.

Data leakage: rigorous time-based splits; ensure future reviews are not leaked into past conditions.

Misalignment with recommendation framing: we narrowed scope to sales forecasting; any recommendation evaluation will be derived from predicted top-K demand changes, time permitting.

S

chieve statistically significant improvement in MAE/sMAPE over sales-only deep baseline on 7-day horizon across a majority of SKUs.

emonstrate that synthetic augmentation yields consistent gains in low-data or high-volatility slices.

rovide qualitative and quantitative evidence that negative sentiment spikes precede sales declines (lead-lag analysis), enabling actionable early warnings.

R