Universiteit Leiden

Comparing local and global forecasting models on highly intermittent time series in manufacturing

Kamil Piotr Mścisz (s4090934)

Supervisors:
Prof. Dr. Joost Visser, Dr. Bahar Rezaei & Dr. Natalia Amat Lefort

Bachelor Thesis – Data Science and Artificial Intelligence

Leiden Institute of Advanced Computer Science (LIACS)
liacs.leidenuniv.nl

July 23, 2026

Abstract

In many production settings, demand forecasting is crucial for efficient resource allocation and a prerequisite for long-term decision-making. Global forecasting models, utilising the whole dataset to train one model across many series, have outperformed local per-series models in recent forecasting competitions. The evidence comes largely from hierarchical retail data with strong cross-series structure. Whether this advantage survives on highly sparse, non-hierarchical data from production shipments largely remains an open question. This thesis asks whether global models remain superior to local ones on moderate-length, highly intermittent demand with no hierarchical structure. The global versus local model comparison is evaluated on an industrial demand dataset: 1,200 weekly SKU series from an industrial label producer. Two global models (DeepAR and LightGBM) are put against three purpose-built local models (Croston's method, TSB, and ADIDA), a local ablation of LightGBM, and three trivial baselines. The evaluation consists of a three-fold rolling-origin cross-validation across metrics spanning two grains and two demand weightings. The answer to the research question is affirmative but narrow: LightGBM beats all the local models on the primary metric, but DeepAR failed to fit, which limits the global representation to a single model. Three qualifications bind this result: no model beats forecasting zero on the weekly grain; equal weighting of all SKUs narrows the advantage of LightGBM; and while forecasting-model ranking remains stable across folds on the primary metric, the overall winner does not. The sparsity of the dataset makes the choice of evaluation grain and weighting as important in shaping the results as the choice of model. The findings suggest LightGBM can be worth considering over local models, and can be a useful model for manufacturers requiring horizon-grain predictions on a highly sparse broad dataset.

Contents

1 Introduction

In many industrial production settings, planning future demand is crucial for efficient operation. Accurate forecasting can prevent stock-outs and reduce delays in order fulfilment. Forecasting demand is a difficult task for several distinct reasons: First, many products are shipped in batches, leaving long runs of zeros, known as intermittent demand, that are challenging for many standard forecasting methods. Second, due to changes in recording procedures, the usable history per product can be short, as older records do not match the current ones, limiting how far back the usable record can go. At the same time, factories tend to produce a large number of products, for all of which a demand time series can in principle be obtained. This often results in datasets with limited depth but extensive breadth, with thousands of short series running in parallel.

Demand prediction models can be split into two groups with different design principles: local and global models. Local models fit a separate model to each time series. Because of this, they are often limited by the amount of data available and can be prone to overfitting, which often forces the choice of a simpler model. Global models leverage the whole dataset by training one model to predict all the time series available. Montero-Manso and Hyndman[1] have shown that global models can match the performance of per-series models, regardless of the series being related. Thanks to pooling the data, the global model can fit a more flexible shared function than any local model can afford, allowing it to compete with the generality of per-series modelling.

An extensive body of research exists comparing local and global models, but most of that literature is limited to retail settings. This thesis provides that comparison on a dataset of weekly demand data for 1,200 product series from MCC Label, an industrial label producer; it compares local (Croston's method, TSB, and ADIDA) and global (LightGBM and DeepAR) models. This comparison addresses the following research question:

RQ: *Do global forecasting models outperform local models on highly intermittent time series in manufacturing?*

1.1 Thesis overview

This section contains the introduction. Section 2 provides the necessary background. Section 3 discusses related work. Section 4 describes the data and methods. Section 5 presents the experiments and their results, Section 6 contains the discussion and limitations, and Section 7 concludes the thesis.

2 Background

2.1 Intermittent Demand

Intermittent demand is characterised by long periods of no demand, interspersed with occasional non-zero requirements. This is common for many production environments, including spare parts production and other industries where products are ordered in infrequent batches. Zero-demand periods can distort smoothing-based forecasts and cause issues with demand-magnitude-normalised metrics due to division by zero. Intermittency poses problems for standard forecasting methods

designed for series with demand in every period, and motivates the use of methods suited to intermittent demand, as described in Section 4.2. In the remainder of this subsection, and whenever intermittent is used together with lumpy, an alternative definition of the word "intermittent" is used to match the phrasing of Syntetos, Boylan and Croston[2]. Elsewhere in this thesis, the wider definition outlined above applies, unless specified otherwise.

A common way to divide sporadic demand patterns is to use a grid along two dimensions representing the distance between non-zero demand and the difference in the scale of the non-zero demand. Syntetos, Boylan, and Croston propose representing those dimensions as, respectively, the mean number of review periods between consecutive non-zero demand timesteps (p) and the squared coefficient of variation of the non-zero demand sizes (CV^2). The reason why their approach was picked is the fact that the cutoff values for the grid were not chosen arbitrarily but rather were derived based on calculations of the mean squared error of two forecasting methods: the original Croston's method and the Syntetos and Boylan method, with the latter performing better on the more intermittent and lumpy series. The exact cutoff values obtained are $p = 1.32$ and $CV^2 = 0.49$, with the grid being comprised of four categories: smooth demand (low p, low CV^2), erratic demand (low p, high CV^2), intermittent demand (high p, low CV^2), and lumpy demand (high p, high CV^2). In that same paper, Syntetos, Boylan, and Croston validated the scheme on a dataset of 3,000 real intermittent-demand series.

2.2 Local and Global Forecasting

This thesis explores two paradigms of time series forecasting: global and local, as outlined by Montero-Manso and Hyndman[1]. The paradigms differ in the number of forecasting models that are fitted to the dataset at hand. The local approach fits one function per series in the dataset, treating each series separately. This approach allows full flexibility in adjusting the function to the series it approximates, but it is often limited by data scarcity, especially in datasets with shorter time series. An alternative approach is to fit a global model to the whole dataset, sacrificing the per-series flexibility to instead learn one function, leveraging the increased data availability to limit overfitting and allow for higher model capacity. This global versus local framing of model comparison is by no means the only one possible. Januschowski et al. propose several ways of classifying models in their paper discussing alternatives for a commonly used split between Machine Learning versus Statistical models[3]. They argue that the divide between ML and Statistical models is mostly tribal in origin, reflecting more the divide between researchers in the forecasting community rather than actual meaningful differences in the models. For that reason, in this thesis the ML versus Statistical model divide will not be used, and the question is framed around the more concrete global versus local divide. The only exception to this rule is when the thesis discusses previous literature that uses the ML versus Statistical divide, in which case the divide is kept for the sake of consistency. Januschowski et al. also note that the distinction is only concerned with how the model parameters are estimated and makes no assumption about whether the underlying series are independent or not. This point lies in line with the equivalence argument we will discuss next.

Montero-Manso and Hyndman[1] show that the loss of generality for global models is not a real restriction. They prove that for any set of per-series functions produced by a local model, there exists a single global function reproducing the same forecasts. The core argument relies on the fact that since the only input to the forecasting method is the series itself with no arbitrary target

attached, a local model fitting a function will predict the same value for two identical series. This means that a situation where the global function would have to predict two different values for two identical series never arises, and thus the global model loses no generality compared to the local one. The argument does not rely on any similarity assumption between the series in the dataset; the series can be drawn from different processes and completely unrelated for it to still hold. This finding leads to the conclusion that a global model can, in theory, match the forecasting of local methods, without the cost of fitting hundreds or possibly thousands of different models, no matter the dataset characteristics.

While global models can represent what local models can, that does not necessarily imply that the global models will generalise better. Montero-Manso and Hyndman address this through a generalisation-bound analysis (their Proposition 2). They find that while the local approach has a better worst-case bound for any series in isolation within the same hypothesis class, when evaluated globally on the dataset the global approach can generalise better if the breadth of the dataset is sufficiently large compared to the length of each series, thanks to the global function not increasing in complexity with dataset breadth as opposed to fitting one function per series, which scales linearly. The core assumption of Proposition 2 is the independence of the series.

In this thesis, we will compare several models from both the global and local paradigms. The local side is represented by three models specifically designed for intermittent demand: Croston's method[4], TSB[5], and ADIDA[6], fitted independently for each SKU series. The global side is represented by two general-purpose models that have been adjusted to accommodate the intermittent nature of the dataset: a gradient-boosted regression-tree model (LightGBM[7]) and a recurrent deep-learning neural network (DeepAR[8]). A detailed description of each model can be found in Section 4.2.

3 Related Work

The M competitions have gradually built a case for the importance of cross-learning in ML-based forecasting methods. Previous literature has found that local ML methods are outperformed by statistical benchmarks[9]. The M4[10] competition has shown a new promising direction for ML methods that could outperform pure statistical methods. M4 was the largest competition of this type to date and utilised a dataset of 100,000 series across six frequencies. For the first time in the M competition history, the three top-performing methods were utilising information from multiple series for prediction. While they were not pure ML methods but rather statistical-ML hybrids, the authors attributed the poor performance of the pure ML entries in the competition to the limited use of cross-learning. They flagged the application of ML methods to forecasting as a direction for future research, with cross-learning as a potential way to improve beyond what the statistical methods can offer.

The M5 competition[11] has provided extensive evidence supporting the use of global models for intermittent demand prediction. The dataset used in the competition comprised sales data for over 3,000 products sold by Walmart in different locations across the US. The dataset consisted of 42,840 series, with the 30,490 lowest cross-sectional aggregation level series, for which the predictions were required, described as intermittent and erratic. For the first time in the M competition history, all of the top 50 methods were ML methods and managed to outperform every benchmark, with LightGBM dominating the top submissions. What is more relevant to us than the ML label is that

all 50 of those submissions utilised cross-learning. The organisers of the competition recognise the dataset structure as a possible reason for the success of global models. Unlike earlier M competitions, M5 utilised a dataset with series being aligned and highly correlated within a retail hierarchy, which could have made cross-learning easier to apply and led to superior results compared to the local approach. The paper flags, as a direction for future research, the need to verify the findings on datasets from outside hierarchical retail data. This thesis contributes to closing this gap by comparing global and local models on non-hierarchical B2B factory shipment data that shares intermittent and erratic characteristics, facilitating the use of similar modelling approaches. One of the important findings of the M5 competition is that the improvement of the winning submission (also LightGBM-based) over Croston falls significantly at the granular end of the hierarchy, where the 77.9% top-level improvement diminishes to single-digit percentages. Our dataset consists of per-product series for a single industrial client, making it the least favourable setting for the global models and facilitating a meaningful test of their advantage.

The closest prior paper to this thesis is Spiliotis et al. 2022[12], which compared 18 models:

- 11 statistical models (including Croston, SBA, TSB, ADIDA, and iMAPA)
- 7 ML models (including Gradient Boosted Trees, RF, MLP, and BNN) – trained in two versions, local and global

The models were trained on a dataset of 3,300 daily SKUs from a Greek retailer. They evaluated the models on two metrics: RMSSE and AMSE, with the results mostly agreeing between the two. The results show that all global models, except for NNs, perform worse than their local counterparts. The most accurate models overall on both metrics were the local versions of GBT and RF, with GBT ahead of RF. The dataset's weak intermittency limits the applicability of this study specifically to intermittent demand. 78% of the series were smooth or erratic under Syntetos, Boylan, and Croston's categorisation scheme* outlined in Section 2.1, leaving only 710 series in the intermittent and lumpy categories. The authors flag that this characteristic could have prevented the intermittent-demand-specific models from showing their full potential. The results provided for intermittent and lumpy series show only the best-performing version of each ML model, which does not allow for direct global versus local comparisons. What we can see is that on these series, SBA achieves statistically indistinguishable performance from the top-two methods (local GBT and local RF), and in general, the gap between top-performing local tree-based methods and other methods (including the two global NN models) gets smaller, compared to the one on the full dataset.

Taken together, these studies leave a specific gap. There are two conflicting views on the local and global comparison. The M competitions clearly point towards the global approach offering better results in comparison with local methods. M5 specifically offers significant evidence in favour of global models, specifically in intermittent demand, based on their hierarchical retail dataset, with strong cross-series structure credited by the authors. On the other hand, Spiliotis et al. show local ML models outperforming their global counterparts (except for NNs) on their weakly intermittent dataset, which they flag as possibly handicapping the intermittency-focused methods, with the lead decreasing on the more intermittent series. They also show that, specifically for their intermittent series, the top ML methods offer statistically the same results as the intermittency-dedicated

*Spiliotis et al. use slightly rounded cutoffs of $CV^2 = 0.5$ instead of 0.49 and $p = 4/3$ instead of 1.32; this difference is almost certainly not meaningful for our comparison.

statistical benchmarks they used, with the gap between global NN methods and top local ML methods shrinking.

The abovementioned literature frames the comparison as statistical versus ML, which can be useful for explaining how data sparsity affects each method, as the methods traditionally thought of as ML tend to be more data-hungry than the statistical ones. That being said, as explained in Section 2.2, Januschowski et al.[3] urge avoiding the statistical versus ML framing as a mostly tribal and fuzzy divide. This thesis explicitly frames the global versus local divide as the primary interest. It is only interested in the traditional statistical versus ML divide with regard to how we can expect the models to behave under data sparsity, and when reporting previous papers that used it. This thesis offers a comparison of global and local models on a purely intermittent/lumpy dataset under the categorisation[2] outlined in Section 2.1. This setting has not been isolated by any one of the papers cited above and is deliberately double-edged: the high intermittency should offer a favourable ground for the less data-hungry local methods (Croston, TSB, ADIDA), while the broad dataset of short series should amplify the benefits of cross-learning, favouring global methods (DeepAR and LightGBM). Next to the sparse-data-specific local methods, we have also included a local version of LightGBM to offer a direct apples-to-apples comparison between global and local models, as Spiliotis et al. did for their ML methods. This leaves us with three model groups, allowing us to isolate the data requirements of the models and the global versus local divide, and compare the data-hungry global, data-hungry local (LightGBM), and less data-hungry local (Croston, TSB, ADIDA) model groups separately.

4 Methodology

4.1 Dataset Overview

The dataset has been obtained in cooperation with MCC Label, a multinational label production company. It comprises 2,386 time series, one per SKU, spanning 126 weeks. The depth of the dataset is limited by changes in recording formats within the company, which constrain the starting point of the usable record.

The models have been trained on a series of realised shipments, indexed by date; shipments are pooled into weekly bins to match the company's existing forecasting process. The start of week 1 corresponds to 2024-01-04, with each week starting on Thursday. Each series starts at the first recorded shipment and runs to the end of the record. The length of the series varies from 1 to 126 weeks, with a median of 94, and only 30 of the 2,386 series run for the full observed period of 126 weeks. The variability in series length is only caused by the different starting points; no series end early. Weeks of no demand have been filled with zeros.

The series have been put through two eligibility filters. The first one filters out the series whose history is too short to meaningfully fit a model (creating the eligible dataset). This threshold has been set at 66 weeks, as it is the shortest period that contains a single 13-week holdout window, a whole year of history, and one more week, containing at least one period that could display yearly seasonality if one were present. The second eligibility filter adds another requirement for the series to be trainable on all three evaluation folds (the evaluation protocol is explained in Section 4.3), and fixes the same set for all folds to make the folds comparable. This increases the required

length to 92 weeks[†], leaving us with a 1,200-SKU subset of the 1,669 SKUs that pass the first filter (filtered dataset). At the earliest origin, the remaining SKUs carry between 53 and 87 weeks of training data, with a median length of 82 (all three statistics are 13 and 26 weeks longer for folds 2 and 3, respectively, reflecting the common endpoint of all series).

Based on the classification from Section 2.1 proposed by Syntetos, Boylan, and Croston[2], the 1,200-SKU dataset on which the models are evaluated comprises exclusively intermittent and lumpy series. The filtered dataset is predominantly intermittent (992 out of 1,200 SKUs).

	p ≤ 1.32	**p > 1.32**
CV² > 0.49	Erratic 0	Lumpy 208
CV² ≤ 0.49	Smooth 0	Intermittent 992

Figure 1: Classification of the 1,200-SKU filtered dataset into categories: erratic, smooth, lumpy and intermittent. Boundary values: $CV^2 = 0.49, p = 1.32$

Apart from the abovementioned classification, a purely informative split at 90% zero-share for the whole series length was introduced to allow a look at the sparsest of the SKUs. This split was chosen arbitrarily, and thus no conclusions should be based on it. Of the 1,200 SKUs passing both filters, 67.5% fall above this threshold. Figure 2 shows the distribution of zero-rate across the filtered dataset, with 1,197 of the 1,200 series having above 50% zero-rate and the mean zero-rate equal to 91.4%. The mean zero-rate across all series is 86.8%, so the filtering leads to a small concentration of the zero-heavy series.

While individual series can show statistically significant autocorrelation, no consistent pattern has been found in how this autocorrelation shows up in different series. When averaged across the filtered dataset, no statistically significant autocorrelation is visible, including on the annual lag of 52 weeks. Figures 3 and 4 show the averaged ACF graph over the 1,200 SKUs evaluated and the ACF graph for nine random series from the filtered dataset. Week 18 has been excluded from the ACF plots because of a non-recurring week-long holiday closure.

[†]A series trainable at the earliest origin needs at least 53 training weeks at said origin. Since all series run to the end of the record, this effectively raises the minimum length to $53 + 13 \times 3$, for the three 13-week-long folds.

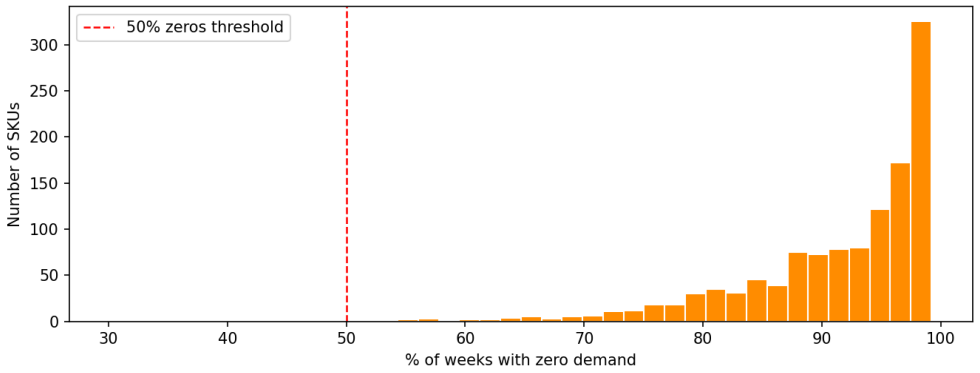

Figure 2: Distribution of the zero-rate in the 1,200-SKU filtered dataset.

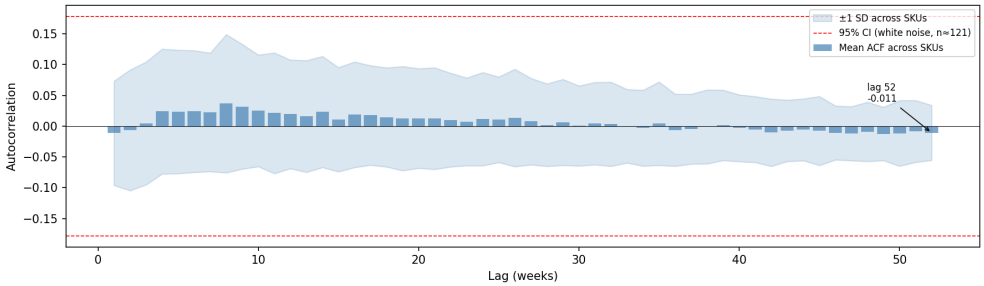

Figure 3: The Auto-Correlation Function (ACF) graph for the whole 1,200-SKU filtered dataset.

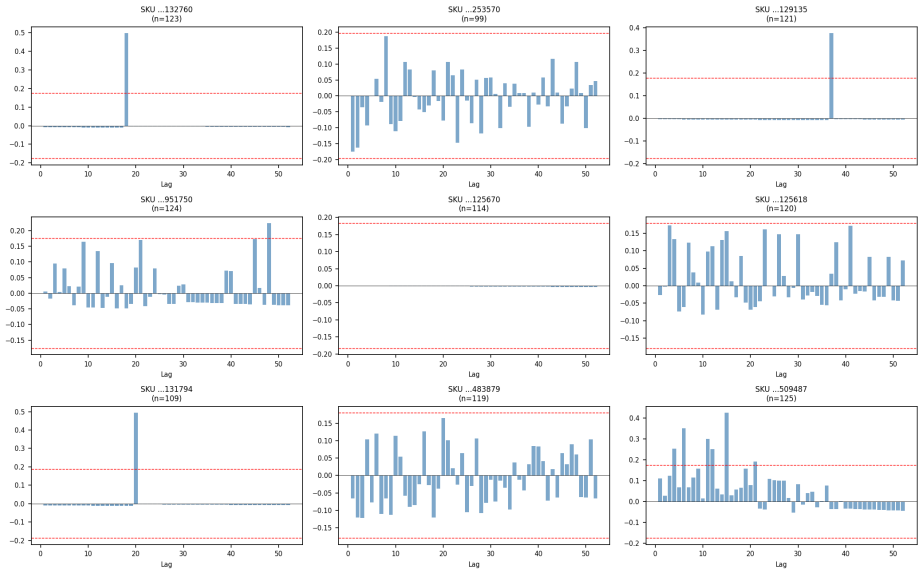

Figure 4: The Auto-Correlation Function (ACF) graph for nine individual randomly selected SKUs out of the 1,200-SKU filtered dataset.

The dataset also exhibits highly variable demand magnitude, varying across several orders of magnitude between SKUs. The distribution of demand magnitude for the 1,200-SKU filtered dataset can be seen in Figure 5 (due to high variability, the demand is presented on a log10 scale).

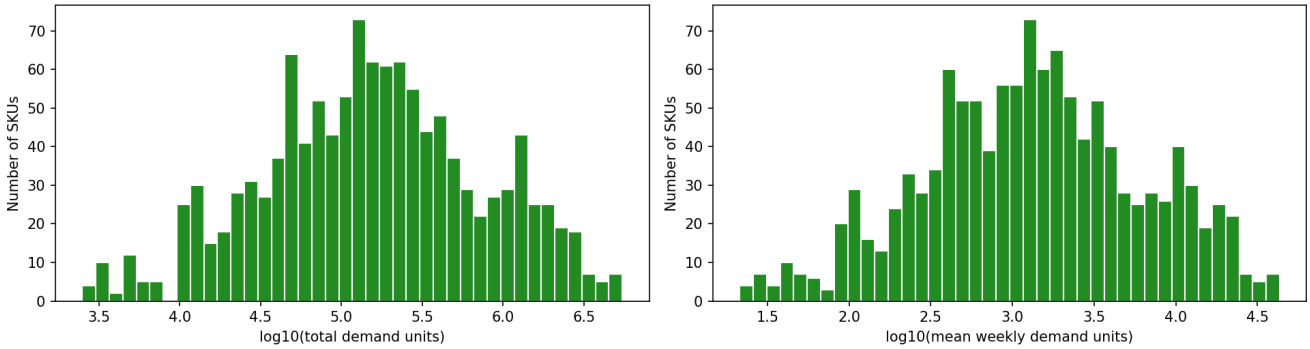

Figure 5: Left: Total demand per SKU in the filtered dataset (log10 scale). Right: Mean demand per SKU in the filtered dataset (log10 scale).

These characteristics were taken into account in the model choice. More conventional local forecasting methods were considered out of the scope of this paper, and instead, methods developed specifically for intermittent demand were utilised. The only models that are not purpose-built for intermittent series among the ones used are the two global models (DeepAR and LightGBM) and the local LightGBM ablation. Those models were adjusted to the dataset through the choice of their likelihood specification (detailed description can be found in Section 4.2.2). Due to the absence of shared seasonality, calendar features were excluded from the models.

4.2 Model Selection

The models evaluated in this paper represent two methodological families: local per-series modelling and global modelling on the whole dataset. The local family consists of three methods specifically designed for intermittent demand, while the global family consists of a gradient-boosted tree model and an autoregressive recurrent neural network. The models have been selected to suit the characteristics of the dataset outlined in Section 4.1.

4.2.1 Local Models

This thesis features four local models, one of which is a local implementation of LightGBM and will be discussed together with the global implementation in Section 4.2.2. The other three methods are established per-series models designed specifically for intermittent series prediction. The three models described in this section have been implemented using StatsForecast[13].

Croston's method[4] decomposes intermittent demand into two factors: non-zero demand occurrence intervals and non-zero demand magnitude, and predicts them separately through simple exponential smoothing. Croston's method only updates its forecasts when demand occurs, which does not let it take into account decay in the series when long periods of no demand are present. This thesis uses the optimised variant from StatsForecast[13], which selects the smoothing parameters by fitting them over a bounded range rather than fixing them, and does so separately for the demand size and interval components.

The Teunter-Syntetos-Babai (TSB) method[5] differs from Croston's method in one key aspect: instead of predicting the intervals between non-zero demand, it predicts the likelihood of non-zero

demand. This seemingly small change carries two benefits over Croston's method. First, non-zero demand likelihood, as opposed to non-zero demand intervals, can be updated in zero-demand periods instead of only when demand occurs. This eliminates the obsolescence problem inherent to Croston's method. Second, Croston's method suffers from a positive bias[14], which the TSB method eliminates[5]. TSB was used with fixed smoothing parameters of 0.1 for both demand size and probability.

A different approach to intermittent time series forecasting is presented by Nikolopoulos et al.[6] in the ADIDA model. ADIDA temporally aggregates the demand into time buckets until intermittency is greatly reduced or completely eliminated. After that, it predicts the demand on the aggregate series and disaggregates the results back to the original frequency.

All three models presented in this subsection are specifically designed for intermittent demand forecasting and produce one fitted model per SKU, placing them firmly on the local side of our comparison.

4.2.2 Global Models

The global models in this thesis are represented by a gradient-boosted tree (GBT) model (LightGBM[7]) and a recurrent neural network (DeepAR[8]).

LightGBM was proposed by Ke et al.[7] as a more efficient solution to fitting decision trees to high-dimensional and large datasets. It utilises Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB) to achieve similar performance to other top GBT models while requiring substantially less time to train. The model has been trained with a Tweedie objective (variance power 1.5). The Tweedie distribution places a positive probability mass at zero while modelling continuous positive values, which makes it well suited to predicting non-negative targets with a large mass at zero[15]. The combination of gradient-boosted trees with the Tweedie distribution has been established as effective in previous literature[16], also being featured among the top contenders of the M5 competition[11]. The feature set consists of lagged demands at lags 1–4, 8, 13, 26 and 52 weeks, rolling means and standard deviations of the one-week lag over 4-, 13- and 26-week windows, and a label-encoded SKU identifier as a static feature. Calendar features were not included because, as per the analysis presented in Section 4.1, the aggregate autocorrelation function shows no seasonality. Negative predictions are clipped to zero, as a negative prediction for demand is nonsensical.

DeepAR[8] is an autoregressive recurrent deep neural network that fits a single set of network weights across all series. DeepAR outputs a parametric conditional distribution at each horizon step. The distribution used for output is the NegativeBinomial, reflecting the overdispersed count nature of the data. The point forecast used for evaluation is the closed-form mean of that distribution, which is also the autoregressive input that is propagated for the following step. The model is conditioned on the previous 52 weeks of each series. This thesis does not analyse the predictive distribution, as all other methods used output a point forecast; a mean point forecast in this case is more suitable for the comparison. This choice means DeepAR cannot predict exactly zero, as the mean of a non-negative distribution will always be strictly positive (unless the whole probability mass is at zero, which is not the case here). This fact carries some consequences for our results, which are presented in Section 5 and discussed in Section 6. It is worth noting that the M5 competition

does not provide direct evidence for DeepAR on intermittent data, but rather flags it as showing potential and worth further investigation[11], to which this thesis aims to contribute.

To provide a direct global versus local comparison isolated from other variables, a local LightGBM variant is included as an ablation. LightGBM-Local trains one LightGBM model per SKU series, with the same hyperparameters and features as the global version, except for the SKU identifier, which carries no information when the model is trained on only one series. A local DeepAR model is not included, because a neural network trained on this sparse per-series data could not achieve any noteworthy performance and the DeepAR architecture is specifically designed to leverage cross-learning.

4.2.3 Trivial Baselines

Alongside the six forecasting models, three trivial baselines were implemented. The Naive baseline predicts the last training-period value for all timesteps forward. ZeroForecast predicts zero for all timesteps forward. HistoricalMean predicts the mean training demand for all timesteps forward. The inclusion of these baselines is crucial for drawing conclusions in this thesis, as with 67.5% of filtered SKUs featuring a zero-rate above 90%, the dataset is extremely sparse, which can prove challenging to more complex models. In particular, ZeroForecast offers an important baseline, because with this many zero-demand periods present, the question of whether the forecasting models can cross the floor set by it becomes relevant and must be answered empirically.

4.3 Evaluation Design

The models are evaluated with a rolling-origin[17][18] three-fold expanding-window cross-validation. The forecast origins are placed 13 weeks apart, producing three non-overlapping 13-week out-of-sample test sets. Every model is retrained per origin on all the data available to that point and forecasts the following 13-week window. This setup allows us to see if the model rankings are time-invariant, or whether different time windows influence different models unequally.

The SKU set used for evaluation is fixed across folds: 1,200 of 1,669 eligible SKUs with at least 53 training weeks available at the first origin point. Thanks to this choice, cross-fold comparisons can be drawn, as with a variable per-fold set, the population variability could influence the comparison. Fixing the set trades 469 SKUs for clean cross-fold comparability. Because of the smaller set, the results obtained in this evaluation are not numerically comparable to those computed on the full eligible dataset.

Scaled metrics are computed strictly per fold: the scaling denominators and weights (the naive-error scales of MASE and RMSSE, the demand-share weights of WRMSSE, and the training-demand rates of the training-scaled WAPE defined below) are recomputed for every fold from the training data available at that fold, to avoid between-fold information leakage. The leaderboard values reported are means of per-fold values unless specified otherwise.

4.4 Metrics

The primary evaluation metric for this thesis is WRMSSE, the volume-weighted root mean squared scaled error. The choice of metric follows the M5 competition[11], with one caveat: due to a lack

Table 1: WAPE design choice grid. Micro metrics pool over all SKU-weeks, while macro metrics pool the per-SKU averages computed over all the weeks.

	Micro (volume-weighted)	**Macro** (equal per SKU)
Weekly grain	*omitted:* rank-identical to MAE	**Training-scaled WAPE** (reported)
Horizon grain	**Horizon-aggregated WAPE** (reported)	*omitted:* out of scope

of data on the unit prices of different SKUs, the value weighting used in the M5 competition had to be replaced with volume weighting. For every SKU, the squared forecast error is divided by the SKU's mean squared naive training-period error. The scaled errors are weighted by the SKU's share of training-period unit demand and combined. The scaling accounts for different SKU magnitudes, while the weighting allows us to still put more emphasis on the high-volume SKUs, which often matter more operationally. The scaled-error principle has been introduced by Hyndman and Koehler[19] as the mean absolute scaled error (MASE).

Alongside WRMSSE, four other metrics are reported: MASE[19], MAE, horizon-aggregated WAPE and training-scaled WAPE. MASE is scaled but not weighted, treating all SKUs as equally important. MAE is not scaled at all, making it more business-legible. The discussion of the two WAPE metrics warrants a separate paragraph later in this section. Those five metrics were picked instead of several others that were considered, for multiple reasons. Weekly-grain micro WAPE and micro RMSSE were computed and rejected because of their mathematical link to MAE and RMSE, respectively. The metrics are essentially rescaled versions of unscaled metrics, making them meaningless for the comparison, as their within-fold rankings necessarily have to be the same as their respective counterparts. For this reason, those variants are not included in this thesis.

WAPE[20] admits two independent design choices. For clarity, the design choice grid can be seen in Table 1. The first choice is aggregation, which determines how SKUs are combined: micro gets heavily biased towards high-volume SKUs and can hide results for the ones with a smaller volume; macro disregards the higher operational value of high-volume items. The second choice is granularity. The different ordering between error calculation and pooling gives us two granularity levels on which the performance is reported by different metrics. Weekly-grain metrics punish wrong forecast placement, where a forecast one week off is punished twice. On the other hand, horizon-grain metrics do not punish misplaced demand within the horizon, focusing on the total in-horizon demand. This allows us to look at whether the models can predict the exact per-week demand, and serves as additional motivation for including two different versions of WAPE. In Section 5, we show that those two ways of looking at prediction accuracy lead to different rankings among the models. The primary comparison of this thesis is on the weekly grain; WRMSSE, MAE, MASE and training-scaled WAPE are all on this grain. The metric chosen for the horizon grain is volume-weighted, because the horizon analysis was introduced to test volume recovery, thereby making horizon-aggregated macro metrics out of scope.

Due to the dataset being highly sparse, a standard macro WAPE version has suffered from structural issues and could not be included in the results. For this version of WAPE, every SKU considered

must have at least one non-zero period in the holdout window; otherwise, the metric for that SKU is undefined. On this dataset, this characteristic results in a silent exclusion of 70% or more of the SKUs within each fold (over 84% for fold 3). This exclusion not only reduces the coverage of this metric but also results in a biased outcome, where the SKUs with full zero holdout periods, which are the most prone to overestimation, get excluded disproportionately compared to the rest. For that reason, a modified version of macro WAPE was included, where the problematic test-set demand value is replaced with training-set demand per SKU (scaled to match the 13-week-long holdout). Since all-zero SKUs were already excluded from the dataset, this construction of the metric keeps all the SKUs in, and is not blind to over-forecasting on dormant SKUs.

Finally, a coarse split between highly sparse SKUs and more active ones was set at an arbitrary value of 90% zero-demand, computed once over each SKU's full observation period. This split is not used to draw any conclusions and exists purely for descriptive convenience. The results computed on this cutoff could prove wrong with a different cutoff and are presented as suggestions for further exploration of highly sparse datasets rather than conclusive evidence.

5 Results

5.1 Overall accuracy

Table 2: WRMSSE mean and per-fold results, and rankings.

model/baseline	WRMSSE				rank		
	mean	fold 1	fold 2	fold 3	fold 1	fold 2	fold 3
ZeroForecast	0.2467	0.3416	0.2309	0.1676	2	1	1
DeepAR	0.2476	0.3418	0.2316	0.1693	3	2	2
LightGBM	0.2587	0.337	0.2567	0.1824	1	3	3
TSB	0.2833	0.361	0.2759	0.2129	4	4	5
ADIDA	0.3028	0.3739	0.3005	0.234	5	6	6
Naive	0.3131	0.4714	0.283	0.1849	8	5	4
LightGBM-Local	0.3624	0.4422	0.3684	0.2765	6	7	7
HistoricalMean	0.4188	0.4646	0.4117	0.3802	7	8	8
Croston	0.6195	0.6688	0.6218	0.568	9	9	9

Table 2 reports a summary of results computed on WRMSSE, including the mean error, errors per fold, and model ranks per fold. The immediate takeaway of the table is that all models, global and local, have failed to consistently outperform the baseline ZeroForecast, which predicts zero demand for every SKU in every week. The only model that manages to beat ZeroForecast on WRMSSE at all is LightGBM, and it does so on fold 1, falling to third place for the other two folds. The second takeaway is that the answer to the research question is affirmative, with full consistency between the folds. Despite some instability in the per-fold rankings, the global LightGBM (mean WRMSSE 0.2587) outperforms TSB (0.2833), ADIDA (0.3028), and Croston (0.6195), not only on the mean WRMSSE but also on every single fold. In the LightGBM versus LightGBM-Local comparison, we can see the isolated effect of cross-learning, as the features, hyperparameters, and objectives were

fixed between the two implementations[‡], with the global version beating the local one on every fold. DeepAR appears second on the mean WRMSSE. In Section 5.2, we show that this placement is not due to the performance of the fitted model, but rather results from a degenerate fit and carries little information about the model as a forecasting method. For that reason, this subsection excludes DeepAR and focuses solely on the comparison between LightGBM and the local models. One caveat visible in this table, worth pointing out, concerns the models' cross-fold performance. The WRMSSE falls monotonically between folds for all nine models and baselines. This reflects the specific features of the evaluation windows, rather than the models themselves. Total holdout demand more than halves across the folds (from 31.1 to 14.8 million units), while the share of zero SKU-weeks rises from 95.2% to 97.4%. This means that the later folds contain substantially less demand to forecast wrongly. For that reason, model performance improvement/deterioration between folds is disregarded in this thesis, as the substantial differences in the holdout window composition make the comparison unfair.

5.2 DeepAR's forecast degeneration

DeepAR failed to learn a per-series forecast on this dataset. The prediction degenerated into a near-zero stable forecast for each SKU across the holdout window. In fold 1, the 846 unique forecasts range from 12.578 to 12.663 units. In fold 2, 15 unique forecasts range from 30.161 to 30.163. In fold 3, the only three unique forecasts can all be rounded to 53.293. DeepAR is the only model with close to no difference in predictions between SKUs within each fold, no matter if the SKU is dormant or shipping hundreds of thousands of units. This leads to the conclusion that DeepAR is not conditioning on the identity or history of the individual series as intended.

The predictions are small enough to make DeepAR functionally indistinguishable from ZeroForecast, but due to the construction of the point forecast as the mean of a non-negative distribution, it remains strictly positive unless all of the distribution mass sits at zero (Section 4.2.2). At the scale of the average weekly demand of 1,427.7 units (computed across all timesteps, including the zero ones) and the median non-zero weekly demand of 26,000 units (excluding zero-demand weeks, as otherwise the median would be 0, which is not informative here), the DeepAR predictions could not just be small predictions of demand; they are indistinguishable from predicting zero. This can also be seen in Table 2, where DeepAR trails right behind ZeroForecast, consistent with being an imperfect, positively biased zero-forecast approximation.

This negative result is important to keep in mind for the remainder of this thesis. The DeepAR figures presented should not be interpreted as reflective of the fitted model's performance but rather as describing a model that failed to learn a proper fit. The low MAE and second-ranked WRMSSE are thus not evidence of the deep-learning model suiting the data or being superior to LightGBM. They are also not evidence of DeepAR's inferiority to ZeroForecast. This conclusion leads us to rest our global versus local comparison on LightGBM. The failure to fit and reliance on LightGBM as the single standing representative of the global models are treated as a limitation of this study (Section 6.3) rather than a property of the DeepAR architecture.

[‡]The SKU identifier is absent from the LightGBM-Local ablation, as in a per-SKU setting it carries no signal and is thus redundant. It does not constitute a meaningful departure from the controlled comparison. A more detailed explanation can be found in Section 4.2.2.

5.3 Ranking stability across folds

Table 3: WRMSSE-based ranking of the forecasting models (without the baselines).

model	rank fold 1	rank fold 2	rank fold 3
DeepAR	2	1	1
LightGBM	1	2	2
TSB	3	3	3
ADIDA	4	4	4
LightGBM-Local	5	5	5
Croston	6	6	6

The apparent ranking instability in Table 2 is caused by inconsistent ranks for the baselines. Table 3 shows the ranking only on the six forecasting models, which remains mostly stable. The only ranking instability within the models is the switch between DeepAR and LightGBM on ranks 1 and 2. Taking into account the degenerate fit of DeepAR discussed in the previous subsection, the switch is a consequence of the ZeroForecast versus LightGBM switch, which happens at the same time, and DeepAR just trails ZeroForecast (with the biggest WRMSSE difference between them being 0.0017). Despite the issues with fitting DeepAR, it is still reported in the table for consistency, but the conclusions drawn are based on the stable rankings between the rest of the models.

The leading cause of the instability of ranking in Table 2 is the Naive baseline climbing from 8th to 5th and 4th between the three folds. This can be explained by the increase in the sparsity of the holdout window as it moves forward. Since Naive forecasts the last value observed, for a dataset this sparse, this value often happens to be zero (respectively for folds 1, 2, and 3: 94%, 96%, and 99% of the time). The number of zero predictions by the Naive forecast increases monotonically, getting closer to ZeroForecast, which is the best-performing baseline of the comparison. Together with the increase in the zero-share of the holdout window (95% for fold 1, 96% for fold 2, and 97% for fold 3), which makes zero predictions more accurate, this leads to a sizable improvement in the rank of the Naive baseline for the later folds.

Two observations follow. First, the global versus local analysis is robust to the choice of forecast origin. The global LightGBM consistently beats all local models and separately beats its own local ablation on every fold and the mean. Second, ranking instability is present only in the baselines (and DeepAR, which is approximately ZeroForecast) and is consistent with the differences in the data in each of the folds mentioned in Section 5.1; relative performance of the baselines shifts as the holdout data changes.

5.4 Horizon grain

The ZeroForecast result obtained on WRMSSE is confirmed by every other weekly-grain metric. ZeroForecast wins on: WRMSSE (0.2467), MASE (0.140), MAE (1,428 units), and training-scaled WAPE (0.230), not only in the average but on every single fold, except for fold 1 on WRMSSE. The ranking among the other models and baselines is not as consistent and will be discussed in the next subsection. As expected in Section 4.4, the weekly-grain double punishment of misplaced

Table 4: Horizon aggregated WAPE results per fold ordered by the cross-fold mean.

model	fold 1	fold 2	fold 3	mean
LightGBM	0.582	0.887	0.839	0.77
ZeroForecast	1	1	1	1
DeepAR	1.003	1.01	1.038	1.017
TSB	0.816	1.09	1.262	1.056
ADIDA	0.95	1.397	1.574	1.307
LightGBM-Local	0.993	1.431	1.653	1.359
Naive	2.138	1.689	1.318	1.715
HistoricalMean	1.641	2.593	3.666	2.633
Croston	4.929	7.464	10.168	7.52

predictions leads to predicting zero being the safe bet. The model that predicts nothing loses only on the week where demand occurs, while a misplaced prediction loses both on the week with demand and on the week where the prediction was placed. In Section 6.1, we will argue that this is as much a statement about the metrics as about the models evaluated. This dominance of ZeroForecast goes away when we change the grain to look at the whole holdout window, delivering the reversal foreshadowed in Section 4.4. Table 4 shows models evaluated on the horizon-aggregated WAPE by fold. By the construction of the metric, ZeroForecast always achieves 1; models with values below 1 explain part of the demand volume left unexplained by predicting nothing. On this grain, models are no longer punished for mistimed predictions, because a prediction placed one week (or several weeks) off contributes the same towards the total predicted demand as a correct one. This difference lets LightGBM pull ahead, with a 23% improvement over ZeroForecast on average (0.77 versus 1). On this metric, LightGBM wins every fold and is the only forecasting model or baseline consistently below the floor set by ZeroForecast. We can also see that the margin by which LightGBM wins substantially increases on the demand-richer fold 1, where it achieves an almost 42% edge over ZeroForecast.

While the superior performance of the global model holds on this metric too, three other models also achieve performance better than ZeroForecast on one of the three folds (fold 1). Those models are TSB (0.816), ADIDA (0.950), and LightGBM-Local (0.993, just beating the 1 of ZeroForecast). DeepAR performs as expected from the WRMSSE results, trailing right behind ZeroForecast, independently confirming the degenerate fit. A model that learned per-SKU demand but misplaced it would see substantial improvements after the grain change, an example of this behaviour being LightGBM. Instead, DeepAR shows very similar results with no visible improvement, leading to the conclusion that neither the magnitude nor the timing of the demand was learned by the neural network. Croston also shows no signs of improvement. It sits far behind all other models and baselines, overestimating the demand.

One qualification has to be stated in this section. Under the descriptive 90% zero-demand split, the superior performance of LightGBM is exclusive to the less sparse section of the data, achieving 0.718 horizon-aggregated WAPE averaged between folds, while on the sparser part of the data, ZeroForecast regains the advantage, with LightGBM scoring 1.454 on average. This is consistent with the mechanism described above: for the most sparse SKUs, there is almost no demand to predict. The predictions cannot be misplaced; part of the positive prediction remains pure error,

no matter the grain. It is important to remember that this split is arbitrary and purely descriptive; these results were not tested for robustness under a different split, and thus no conclusions can be drawn from them. They are only reported to help with the interpretation of other results and point towards questions that were not answered in this thesis and could be explored in further research (Section 6.4).

5.5 Metric disagreement and the core finding

The tension outlined in Section 4.4 between the micro and macro metrics, where the former are biased towards high-volume SKUs, while the latter completely disregard their operational weight, is not just theoretical on this dataset. The two metric groups do not fully agree on the central comparison of this thesis.

Table 5: Cross-fold mean results for the weekly-grain metrics.

model	WRMSSE	MAE	MASE	training-scaled WAPE macro
ZeroForecast	0.2467	1428	0.14	0.23
DeepAR	0.2476	1458	0.194	0.295
LightGBM	0.2587	2378	0.281	0.444
TSB	0.2833	3029	0.282	0.477
ADIDA	0.3028	3450	0.347	0.593
Naive	0.3131	2697	0.262	0.434
LightGBM-Local	0.3624	3342	0.339	0.552
HistoricalMean	0.4188	5194	0.745	1.158
Croston	0.6195	11572	13.57	15.025

Table 6: Training-scaled WAPE results per fold ordered by the cross-fold mean.

model/baseline	training-scaled WAPE (macro)				rank		
	mean	fold 1	fold 2	fold 3	fold 1	fold 2	fold 3
ZeroForecast	0.2296	0.2473	0.2774	0.1641	1	1	1
DeepAR	0.2947	0.2684	0.3355	0.2802	2	2	3
Naive	0.434	0.5388	0.5351	0.2282	4	4	2
LightGBM	0.4437	0.4319	0.5432	0.3561	3	5	4
TSB	0.4767	0.5419	0.5221	0.3661	5	3	5
LightGBM-Local	0.5521	0.6401	0.6032	0.4129	6	6	6
ADIDA	0.5933	0.6434	0.6423	0.4943	7	7	7
HistoricalMean	1.1585	1.1549	1.2076	1.1129	8	8	8
Croston	15.0251	13.0086	15.1392	16.9274	9	9	9

Table 5 presents the cross-fold average results for all the weekly-grain metrics. The volume-dominated metrics broadly agree with each other and with Section 5.1. WRMSSE, MAE, and horizon-grain WAPE all agree that LightGBM is the best-performing forecasting model among the non-degenerate ones. While on average between folds LightGBM scores the best on both training-scaled WAPE (0.4437 versus 0.4767 for TSB, the best-performing local model) and MASE

Table 7: MASE results per fold ordered by the cross-fold mean.

model/baseline	MASE				rank		
	mean	fold 1	fold 2	fold 3	fold 1	fold 2	fold 3
ZeroForecast	0.1396	0.1417	0.1864	0.0907	1	1	1
DeepAR	0.1941	0.1592	0.2351	0.188	2	2	3
Naive	0.2618	0.3135	0.3443	0.1276	4	4	2
LightGBM	0.2812	0.2562	0.3643	0.223	3	5	5
TSB	0.2824	0.3138	0.3276	0.2058	5	3	4
LightGBM-Local	0.3392	0.3938	0.3852	0.2386	7	6	6
ADIDA	0.3472	0.3684	0.3947	0.2784	6	7	7
HistoricalMean	0.7448	0.7352	0.7929	0.7062	8	8	8
Croston	13.5699	11.6747	13.6667	15.3684	9	9	9

(0.2812 versus 0.2824 also for TSB), the win is not nearly as clear as for the volume-weighted metrics, and for MASE it is negligible, although still present. When we look at separate folds, we can see that on training-scaled WAPE, TSB beats LightGBM on fold 2 (Table 6), and on MASE, TSB beats LightGBM on folds 2 and 3 (Table 7). This shows that, as opposed to the volume-weighted metrics, the superiority of the global method over the top-performing local model is not stable across folds.

On both macro metrics, all the models are outperformed by ZeroForecast on every fold. The Naive baseline beats all the models (not counting DeepAR) on average.

This has a direct consequence for the headline answer of this thesis. The answer is clearly affirmative on WRMSSE, the primary metric, with the horizon-grain WAPE and MAE additionally supporting this finding. The macro metrics do not change that answer; on the cross-fold mean, LightGBM remains better than all the local models; that being said, the evidence on macro metrics is not as one-sided: LightGBM ranks above TSB in only three out of six fold-by-metric comparisons. Reporting only the volume-weighted micro metrics would overstate what this thesis has established. The headline answer is therefore: the global model outperforms the local models on the cross-fold mean, but the lead diminishes substantially on MASE and less strongly on training-scaled WAPE – although it does not go away. The mechanism behind this is discussed in Section 6.1. What actually changes under the macro metrics is the placement of the Naive baseline. The same mechanism outlined in Section 5.3, which explained the instability of this baseline, now explains its increased relative performance. Since the Naive baseline overwhelmingly predicts 0, it performs better on the sparsest of the SKUs, which count less towards the micro metrics because they account for less volume. On the macro metrics, the volume no longer plays a role, so the relative performance of the Naive baseline improves to rank 3, a visible improvement over the 6th on the mean WRMSSE.

5.6 Croston's method

Croston's method is the worst-performing model on every metric by a wide margin. On WRMSSE, the error increases compared to the next worst one by 48%, on MAE the error more than doubles, and on horizon-grain WAPE, it almost triples. The biggest margin can be observed for the macro metrics, where Croston achieves an error over 18 times (MASE) and almost 13 times (training-scaled

WAPE) as large as the next-worst HistoricalMean baseline.

This behaviour is not unexpected. Croston's method does not update its predictions between periods of non-zero demand[4], which leads to it performing worse on SKUs whose demand falls, especially ones that go dormant. When an SKU becomes dormant, Croston cannot update its prediction and continues to predict outdated values from the last non-zero week update. This is the exact issue TSB[5] solves by using a non-zero-demand probability updated every period, including the zero-demand ones, which allows for prediction decay in long periods of no demand.

On this dataset, dormant SKUs and long periods of no demand are not edge cases, but the bulk of the series. We can see this directly in the backtest. Because of the moving window backtest setup, we can see that for 722 SKUs, which have zero demand in all of the holdout windows, predictions for all three folds are not updated and remain identical. The level at which the forecast for those dormant SKUs freezes is not small, with a median of 5,801 units per week. With the actual demand being zero, this blows up the errors for Croston. Across all of the folds, 90.6% of Croston's total absolute error is incurred on zero-demand SKU-weeks, compared to 49.0% for LightGBM, 62.8% for TSB and 67.7% for ADIDA.

In practice, this shows why rejecting the conventional macro WAPE from Section 4.4 was warranted. The metric is undefined precisely for those zero-holdout-window-demand SKUs, from which the overwhelming majority of Croston's error comes. On that metric, Croston's complete failure on dormant SKUs would not be visible. Dropping the zero test-demand SKUs leads to the conventional WAPE macro error of Croston being higher than the next-worst HistoricalMean by a factor of 1.3, instead of the factor of 13 on the training-scaled WAPE, which does not drop any SKUs.

6 Discussion and future work

6.1 ZeroForecast floor, grain, and weighting

Section 5.4 established that the ZeroForecast baseline tops all of the weekly-grain metrics, with the exception of fold 1 on WRMSSE, but does not keep that lead when the grain is expanded to the whole holdout window. This divide is caused by the sparsity of the dataset. On a dataset with over 90% zero-demand weeks, predicting zero all the time is correct in an overwhelming majority of the weeks, while predicting demand risks mis-timing the prediction and running into the double-punishment mechanism mentioned in Section 4.4. ZeroForecast already predicts the correct demand for more than nine out of ten SKU-weeks, since demand in those weeks is zero, so a non-zero forecast does not have much accuracy to gain on the weekly grain. On the other hand, at the horizon grain, predicting zero demand leaves all of the non-zero demand unexplained, which makes ZeroForecast beatable, as even mis-timed predictions can win on this grain. The horizon grain allows the metric to distinguish accurate models from silent ones. This means that, regardless of our global versus local comparison, none of the models are better than predicting nothing for the weekly-grain accuracy, but LightGBM demonstrably recovers demand volume over the full horizon, with three other models also managing to beat forecasting all zeros on one of the three folds.

The grain discussion cannot be solved on this data; it is a decision every application of demand forecasting has to make for itself. The choice between micro and macro metrics is similarly not

one that the data can settle, but rather a purpose decision. Whether the errors on high-volume SKUs count for more than the errors on the low-volume SKUs has to be decided on an application basis. As noted in Section 4.4, the macro metrics were only computed on the weekly grain, meaning that the comparison has to be done exclusively on said grain. The weekly grain is exactly the place where no models can beat ZeroForecast, and the macro metric comparison inherits this floor. Removing the volume weighting favours the sparser series, making ZeroForecast even harder to beat, and allowing the Naive baseline to also get ahead of the forecasting models (excluding DeepAR). Despite the smaller lead of LightGBM, the global versus local answer does not flip. That being said, the diminished lead weakens the affirmative under the macro metrics.

6.2 What survives

The previous section's qualification leaves us with three surviving findings, on which the contribution of this thesis rests.

The first is the headline answer to the research question. Because DeepAR failed to fit (Section 5.2), the global models are represented by LightGBM alone, making it a comparison not between global models as a class, but between a specific established[11] global model and local models. The claim is not that global models outperform local ones, but rather that LightGBM specifically beats the local models evaluated in this thesis, and it does so consistently. On the primary metric (WRMSSE), LightGBM beats all the designated local models and its local ablation on every single fold and on the cross-fold average. The victory of the global LightGBM over the local ablation specifically isolates the cause. Since the local and global versions share the design, the gain in accuracy can be attributed specifically to cross-learning, and not to the model class.

The second is that the LightGBM horizon-grain result is not just a metric artefact. ZeroForecast always scores exactly 1 on the horizon-aggregated WAPE by construction. A better performance means that the model explains part of the demand left unexplained by predicting zero. LightGBM scores below 1 on every fold, meaning it consistently reflects demand volume recovered over the lead time. This property is completely concealed on the weekly-grain metrics. The exploratory split points towards this explanatory power of LightGBM being concentrated on the less sparse series (0.718 average horizon-aggregated WAPE on series with below 90% zero rate). The horizon-aggregated WAPE of LightGBM on the more than 90% zero-rate series increases to 1.454, above ZeroForecast by a sizable margin. Of course, we do not know if this discrepancy survives moving the split, and the cutoff point is arbitrary, but it does at least point towards LightGBM performing better on less sparse series, where demand volume is concentrated.

The third concerns the evaluation design and stability of the findings. The ranking between the forecasting models (when we disregard DeepAR, as a degenerate fit) remains stable across all folds on WRMSSE. The main global versus local finding would have been the same had we performed the evaluation on a single holdout window. What is unstable is the actual winner. The three-fold evaluation allows us to see this instability and adjust the strength of the LightGBM versus ZeroForecast claim accordingly, whereas with a single holdout window, different origins would yield a different winner. The margins between ZeroForecast and LightGBM are in the order of thousandths of WRMSSE, making them too slim for a single holdout window to reliably select the winner. By making it a three-fold evaluation, we can recover some of that reliability. This is

the methodological payoff of the rolling-origin design specified in Section 4.3, which justifies the exclusion of some SKUs.

6.3 Limitations

Tuning was not performed on either of the global models, a plausible factor for DeepAR's failure to obtain a proper fit. The output of DeepAR was evaluated as a point forecast, and no tests were performed on whether the probability distribution output carries information that was lost by the point statistic. The tuning limitation mostly concerns DeepAR, as LightGBM's performance was already above that of the local models without the tuning, whereas for DeepAR the tuning could facilitate a proper fit.

The three folds constrain what can be claimed: no significance tests were performed; we instead report the per-fold values, means, and ranks. The fixed backtest population of 1,200 out of the 1,669 eligible SKUs trades coverage for cross-fold comparability. The findings do not necessarily extend to the excluded SKUs.

The 90% zero-share split is arbitrary, and no conclusions in this thesis can rely on its placement.

All findings rest on a single dataset of realised shipments from one producer to one client, with all series falling into intermittent or lumpy categories. This level of intermittency is extreme even for the intermittent-demand literature. The findings of this thesis should not be extended beyond this extreme demand regime. Spiliotis et al.[12] conducted a global versus local model comparison on a weakly intermittent dataset, which showed most local models performing better than their global counterparts (including boosted trees). This thesis is not evidence against their findings because of the differences in the intermittency regime of the two datasets used in the studies. Together, their study and this thesis point towards the global versus local model comparison flipping when dataset characteristics change, with intermittency likely being a factor, although since LightGBM was not one of the models in their study, this cannot be stated as a concrete conclusion.

6.4 Further work

The most direct extension is evaluating both global models with tuning and on their probabilistic outputs, with fixing DeepAR's degenerate fit as a prerequisite. DeepAR natively produces a predictive distribution, and LightGBM admits quantile objectives. Training LightGBM using a quantile objective instead of Tweedie and scoring both models' quantiles with pinball loss would evaluate the outputs most directly relevant for safety-stock quantities.

A second possible extension would involve translating the forecasts into a production-planning simulation and estimating the operational impact of the predictions. The M5 organisers note that the competition forecasts were not linked to operational costs[11].

Finally, the extension pointed towards by the arbitrary 90% zero-rate split is routing. On this split, LightGBM roughly doubles its error on the more sparse SKUs. Routing different SKUs to different models based on their sparsity could not only yield better results but also test whether the observation survives a principled threshold.

7 Conclusion

This thesis conducted a comparison of global and local forecasting models on a weekly industrial demand dataset for 1,200 SKU series from MCC Label, under an entirely intermittent or lumpy demand regime, under the categorisation of Syntetos, Boylan, and Croston[2]. The local models are represented by methods specific to intermittent demand (Croston's method, TSB, ADIDA), while the global models are represented by LightGBM and DeepAR. Additionally, a LightGBM local ablation (LightGBM-Local) is included as a direct counterpart to the global LightGBM. Alongside these forecasting models, three trivial baselines are included: Naive, ZeroForecast, and HistoricalMean. All methods were evaluated under a three-fold rolling-origin cross-validation, on weekly grain weighted by demand, weekly grain unweighted, and horizon grain weighted metrics.

The answer to the research question is affirmative but layered. LightGBM beats all the local models on the primary metric across every fold. It also outperforms the local ablation in every fold, isolating cross-series learning rather than the model class as the source of the performance increase. The comparison rests purely on LightGBM. DeepAR produced a degenerate fit, collapsing into a near-constant forecast trailing ZeroForecast on performance. This result is treated as a limitation rather than a result on the DeepAR architecture and leads to the exclusion of DeepAR from the main comparison. The finding is thus limited to LightGBM on the side of the global models and should be read as such, instead of being generalised to the whole class of global models.

The scope of this answer is limited by three qualifications. First, all models, local and global, fail to beat the ZeroForecast baseline on the weekly grain. The sparsity of the data makes predicting nothing the choice that maximises accuracy, when a mistimed prediction is punished twice. Only when looking at the horizon-aggregated metrics does LightGBM manage to beat ZeroForecast, explaining some of the total demand that predicting zero could not explain. Second, the advantage of LightGBM over the best-performing local model narrows on macro, equal-weighted metrics. While the advantage survives the metric switch, this decline weakens the claim of global model superiority on metrics where all SKUs are treated equally, no matter their volume. Third, while the main-metric ranking between the forecasting models remains stable across folds, the ranking becomes unstable when baselines are also included. This does not influence the headline global versus local comparison, but it does mean that the best-performing method across all forecasting models and baselines is not stable across folds. The three-fold design allows us to isolate this unstable result from the stable findings, whereas a single-fold evaluation would crown a winner on an unstable margin in the thousandths of WRMSSE.

Together, these findings suggest that on a dataset with this extreme intermittency, the choice of evaluation metrics on both grain and weighting matters for the findings as much as the models themselves. The same models can be judged to be worse than predicting nothing or to outperform it, depending on the metric. That being said, the main finding of global LightGBM outperforming all local forecasting models has been stable across the metrics.

The most direct extensions identified in this thesis are probabilistic evaluation of models' output, a production-planning simulation linking accuracy to operational cost, and implementation of sparsity-based routing. Those extensions are left to be pursued by future work.

References

[1] Pablo Montero-Manso and Rob J Hyndman. Principles and algorithms for forecasting groups of time series: Locality and globality. *International Journal of Forecasting*, 37(4):1632–1653, 2021.

[2] Aris A Syntetos, John E Boylan, and John D Croston. On the categorization of demand patterns. *Journal of the Operational Research Society*, 56(5):495–503, 2005.

[3] Tim Januschowski, Jan Gasthaus, Yuyang Wang, David Salinas, Valentin Flunkert, Michael Bohlke-Schneider, and Laurent Callot. Criteria for classifying forecasting methods. *International Journal of Forecasting*, 36(1):167–177, 2020.

[4] John D Croston. Forecasting and stock control for intermittent demands. *Journal of the Operational Research Society*, 23(3):289–303, 1972.

[5] Ruud H Teunter, Aris A Syntetos, and M Zied Babai. Intermittent demand: Linking forecasting to inventory obsolescence. *European Journal of Operational Research*, 214(3):606–615, 2011.

[6] Konstantinos Nikolopoulos, Aris A Syntetos, John E Boylan, Fotios Petropoulos, and Vassilis Assimakopoulos. An aggregate–disaggregate intermittent demand approach (ADIDA) to forecasting: an empirical proposition and analysis. *Journal of the Operational Research Society*, 62(3):544–554, 2011.

[7] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*, 30, 2017.

[8] David Salinas, Valentin Flunkert, Jan Gasthaus, and Tim Januschowski. DeepAR: Probabilistic forecasting with autoregressive recurrent networks. *International Journal of Forecasting*, 36(3): 1181–1191, 2020.

[9] Spyros Makridakis, Evangelos Spiliotis, and Vassilios Assimakopoulos. Statistical and machine learning forecasting methods: Concerns and ways forward. *PloS one*, 13(3):e0194889, 2018.

[10] Spyros Makridakis, Evangelos Spiliotis, and Vassilios Assimakopoulos. The M4 competition: 100,000 time series and 61 forecasting methods. *International Journal of Forecasting*, 36(1): 54–74, 2020.

[11] Spyros Makridakis, Evangelos Spiliotis, and Vassilios Assimakopoulos. The M5 competition: Background, organization, and implementation. *International Journal of Forecasting*, 38(4): 1325–1336, 2022.

[12] Evangelos Spiliotis, Spyros Makridakis, Artemios-Anargyros Semenoglou, and Vassilios Assimakopoulos. Comparison of statistical and machine learning methods for daily SKU demand forecasting. *Operational Research*, 22(3):3037–3061, 2022.

[13] Federico Garza, Max Mergenthaler Canseco, Cristian Challú, and Kin G Olivares. StatsForecast: Lightning fast forecasting with statistical and econometric models. *PyCon Salt Lake City, Utah, US*, 2022:6, 2022.

[14] Aris A Syntetos and John E Boylan. On the bias of intermittent demand estimates. *International Journal of Production Economics*, 71(1-3):457–466, 2001.

[15] Yi Yang, Wei Qian, and Hui Zou. Insurance premium prediction via gradient tree-boosted Tweedie compound Poisson models. *Journal of Business & Economic Statistics*, 36(3):456–470, 2018.

[16] Olivier Sprangers, Wander Wadman, Sebastian Schelter, and Maarten de Rijke. Hierarchical forecasting at scale. *International Journal of Forecasting*, 40(4):1689–1700, 2024.

[17] J Scott Armstrong and Michael C Grohman. A comparative study of methods for long-range market forecasting. *Management Science*, 19(2):211–221, 1972.

[18] Leonard J Tashman. Out-of-sample tests of forecasting accuracy: an analysis and review. *International Journal of Forecasting*, 16(4):437–450, 2000.

[19] Rob J Hyndman and Anne B Koehler. Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22(4):679–688, 2006.

[20] Stephan Kolassa and Wolfgang Schütz. Advantages of the MAD/MEAN ratio over the MAPE. *Foresight: The International Journal of Applied Forecasting*, (6):40–43, 2007.