COMPARATIVE ANALYSIS OF MODERN MACHINE LEARNING MODELS FOR RETAIL SALES FORECASTING

Luka Hobor[1,2], **Mario Brcic**[1,2], **Lidija Polutnik**[3], **Ante Kapetanovic**[4]

[1] University of Zagreb, Faculty of Electrical Engineering and Computing, Unska 3, 10000 Zagreb, Croatia
[2] It From Bit d.o.o., Zagreb, Croatia
[3] Babson College, Babson Park, Massachusetts, United States
[4] mStart Plus d.o.o., Zagreb, Croatia
{Luka.Hobor,Mario.Brcic}@fer.hr, polutnik@babson.edu, Ante.Kapetanovic@mstart.hr

ABSTRACT

Accurate forecasting is key for all business planning. When estimated sales are too high, brick-and-mortar retailers may incur higher costs due to unsold inventories, higher labor and storage space costs, etc. On the other hand, when forecasts underestimate the level of sales, firms experience lost sales, shortages, and impact on the reputation of the retailer in their relevant market. Accurate forecasting presents a competitive advantage for companies. It facilitates the achievement of revenue and profit goals and execution of pricing strategy and tactics. In this study, we provide an exhaustive assessment of the forecasting models applied to a high-resolution brick-and-mortar retail dataset. Our forecasting framework addresses the problems found in retail environments, including intermittent demand, missing values, and frequent product turnover. We compare tree-based ensembles (such as XGBoost and LightGBM) and state-of-the-art neural network architectures (including N-BEATS, NHITS, and the Temporal Fusion Transformer) across various experimental settings. Our results show that localized modeling strategies especially those using tree-based models on individual groups with nonimputed data, consistently deliver superior forecasting accuracy and computational efficiency. In contrast, neural models benefit from advanced imputation methods, yet still fall short in handling the irregularities typical of physical retail data. These results further practical understanding for model selection in retail environment and highlight the significance of data preprocessing to improve forecast performance.

Keywords Retail Sales Forecasting, Time-Series Analysis, Machine Learning, Deep Learning, Gradient-Boosted Decision Trees, Neural Networks, Predictive Analytics, Retail Analytics

1 Introduction

Accurate sales forecasting is indispensable in retail, enabling better inventory planning, resource allocation, and achieving revenue and profit goals. The choice of the best possible forecasting model to use is of the essence as retailers aim to gain a competitive advantage in the context of irregular demand and limited historical sales data in various product categories. While classical approaches such as ARIMA [1] have historically formed the backbone of retail forecasting, they often fall short in capturing modern retail complexities and operational nuances. Recent advances in machine learning and deep learning present significant opportunities for retailers to improve their sales predictions and operational efficiencies with new alternatives, such as tree-based models like XGBoost [2] and LightGBM [3], as well as neural architectures such as N-BEATS [4], NHITS [5], TFT [6], and others [7, 8, 9, 10, 11].

Recent comparative evaluations of forecasting models have shown mixed results across domains. Studies such as [12, 13, 14, 15, 16, 17] have explored the performance of neural networks versus gradient boosting models and highlighted that deep learning does not consistently outperform tree-based approaches, especially on tabular, sparse, or highly intermittent retail data. These findings align with results from the M5 forecasting competition [18], which used Walmart's highly granular retail sales data consisting of daily item-store records. While some series exhibited

intermittent demand patterns, the dataset primarily featured dense, continuous observations with limited explicit missingness. Despite this, LightGBM-based ensembles outperformed deep learning models, including those developed by Amazon's forecasting team [19]. On the other hand, recent findings from Zalando suggest that transformer-based models exhibit scaling laws in retail forecasting: as the volume of training data increases, demand forecasting error decreases in a predictable manner [20]. These understandings motivate a closer investigation into the conditions under which different forecasting paradigms perform best.

Much of the recent deep learning success has been in large-scale online retail settings, where companies like Amazon [21] and Zalando [20] operate centralized warehouses and enjoy the benefits of aggregated demand and operational homogeneity. In contrast, brick-and-mortar (B&M) retail is far more fragmented: sales occur across thousands of physical stores, each with limited shelf space, store-level variability, and frequent changes in product assortment. These conditions lead to noisy, intermittent demand signals known to affect neural forecasting models[22, 23, 24, 25, 26].

In this study, we conducted a large-scale evaluation of forecasting models in the context of B&M retail, using real-world data from a major South-East Europe (SEE) retailer. The forecasting goal was to provide the best sales prediction possible for daily demand 365 days into the future for products in the hygiene product category based on their daily demand data. The retailer in this study operates on an annual planning cycle. Models deployed by the research team are benchmarked for their ability to handle operational realities of physical retail, including intermittent, missing, and product censorship due to assortment shifts and other changes in the retail environment. Our work builds on lessons from both academic literature and industrial benchmarks, including the Rossmann case study [27] and the M5 competition [18], and provides new insights into evaluation of model performance.

Our main contributions are as follows:

- We describe an end-to-end modeling pipeline for long-horizon retail forecasting.
- We provide empirical results comparing state-of-the-art forecasting models, including statistical, machine learning, and neural approaches, on a real-world, high-resolution B&M retail dataset.
- We highlight the limitations of neural models under operational constraints typical of physical retail and provide insights on how to improve forecasting accuracy in the B&M environment, offering practical guidance for academics studying performance of forecasting models in the retail environment as well as retail practitioners.

2 Dataset

2.1 Data analysis

The dataset includes daily retail sales information with multiple dimensions, including time, store characteristics, product attributes, prices, promotional activities, and inventory changes. Each observation represents a daily record for a specific product-store combination, making the dataset well-suited for longitudinal analysis across multiple hierarchical levels.

Each store is assigned a strategic zone, not geographically defined, but based on business strategy, which determines the price structure of products. Products themselves follow a two-level hierarchy: individual items are first organized into subcategories called "groups," and within each group, related products are further clustered into units-of-need (UoNs). This ordering, groups first, then UoNs, enables precise, granular modeling and targeted price strategies.

Product metadata includes category, group, unit pricing (with and without tax), and cost of goods sold. Promotions are extensively captured through binary and count features indicating tactical promotions and various loyalty initiatives. Inventory data is tracked comprehensively, including physical stock levels, incoming shipments, and reserved or frozen quantities.

The daily time series in this dataset, as can be seen in Table1, display highly irregular temporal patterns, with over 70% classified as intermittent and most of the remainder falling into lumpy or erratic categories, while smooth series are rare. This irregularity, combined with an average coverage ratio of 63% and 50% missingness in the training set, presents challenges for traditional forecasting methods. The series' forecastability classification is done using the Syntetos-Boylan method[28]. The product sales of the company studied can be described by a dynamic environment with shifting demand, changes in competition, and differences in turnover across products resulting in dynamic impacts on sales, and assortment changes in the stores. For example, many product series are censored, observed for only part of the total time window, typically around six months, due to frequent product introductions and discontinuations. A typical example of a censored product would be seasonal items, such as decorative napkins for Easter or Christmas.

Classification	Total Count	Percentage (%)
Erratic	2345	3.11
Intermittent	52816	70.06
Lumpy	17704	23.48
No Demand	692	0.92
Smooth	1830	2.43

Table 1: Series Demand Classification

Metric	Train	Valid
Series Count	70201	54454
Global Average Missingness	0.50	0.30
Global Average Coverage Ratio	0.63	0.82
Eliminated/New Products Count	20933	5186
Eliminated/New Ratio	0.30	0.10
Series with More Than 730 Data Points	6966	—
Imputed Series with More Than 730 Data Points	69442	—

Table 2: Train vs Valid Series Statistics

This limits historical visibility and can bias modeling efforts if not addressed a priori. Around 30% of products were eliminated during training and 10% are newly introduced in validation.

The dataset structure and challenges are comparable to those reported in major demand forecasting benchmarks such as the M5 Competition[18] or Rossman study case [27] and industry-specific datasets from companies like Zalando [20] and Amazon[21]. However, unlike some e-commerce datasets which benefit from smoother, denser demand signals, B&M retail environments such as the one studied here are more exposed to changes in consumer demand, localized promotions, and physical differences across stores, which further drive stability of consumer demand.

Together, these characteristics highlight the need for robust forecasting models that can handle intermittent demand, assortment volatility, and varying data coverage challenges common in real-world retail but often underrepresented in more sanitized datasets in academic research.

2.2 Data Preprocessing

To ensure high-quality inputs for forecasting, a structured data preprocessing pipeline was implemented, including enrichment, imputation, transformation, and formatting steps appropriately adjusted for time series models.

Competitive and Macroeconomic Environment. The point of sale store specific data was augmented with relevant competitive and macroeconomic indicator data. In order to control for the level of competition in the relevant geographic market, the count of competitors within a 1-kilometer radius and their prices for comparable items were added to the dataset. Further, macroeconomic environment variables from the National Statistical Office, such as Consumer Price Index (CPI), average salaries (national and regional), and population estimates within specific geographies, were included in order to control for the purchasing power and the size of the market.

Handling Missing Values. Basic imputation strategies were applied first, including forward and backward fills for time-continuous numeric variables, and mode imputation for categorical fields. These methods were chosen for fields with stable or slowly changing, predictable behavior. In addition, we conducted a separate experiment with a deep learning-based imputation model to assess its impact on downstream predicting accuracy. Specifically, we used the SAITS model [29], implemented via the PyPOTS library [30], to impute missing values in the training and validation sets. This approach aimed to restore temporal consistency in volatile variables and evaluate whether improved continuity would benefit model performance. Figure 1 illustrates a representative example of the imputation process, demonstrating how the applied methods restore temporal consistency in the data.

Feature Engineering. Several engineered features were introduced to enrich the temporal signal. Lag and rolling-window statistics for sales and promotional indicators were created to capture historical dynamics, particularly beneficial for models like XGBoost and LightGBM that treat each row independently. All nominal prices were converted into real prices using CPI, and relative pricing metrics were computed: one measuring the chain's pricing advantage or disadvantage versus competitors (out-store metrics), and another capturing intra-store product positioning within the same UoN.

Feature Selection. Key variables used in the demand forecasting models predicted by the economic theory were retained. To reduce model complexity and improve generalization, we applied a filtering procedure on additional predictors based on feature importance. We used the Boruta algorithm [31] with LightGBM as the background model to identify and keep only features with proven predictive value. Features with excessive missingness, low variance, or multicollinearity were discarded to enhance efficiency and stability.

Train-Validation Split. A time-based cutoff was used to divide the dataset into training and validation periods. For ensemble models, we retained only series with observations in both periods, while for neural models, we further restricted to series having at least 730 training points.

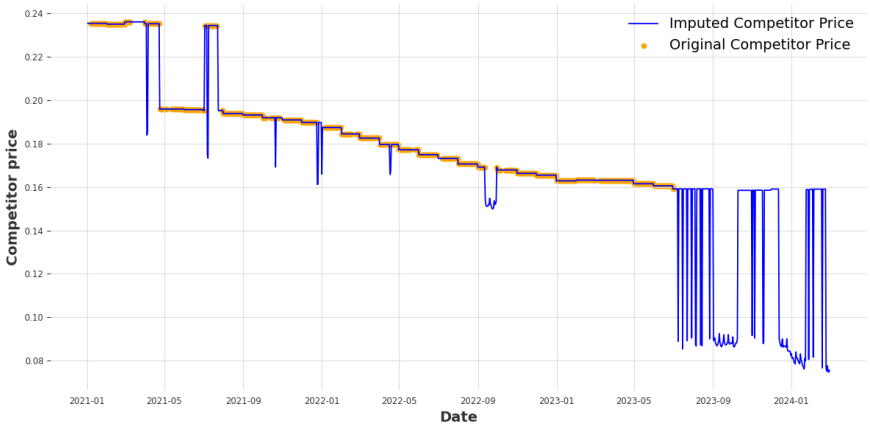


Figure 1: Competitor Price Feature Values Real and Imputed

Categorical Encoding. Categorical variables were encoded using the CatBoost encoder [32] from the package `category_encoders`.

Final Time-Series Transformation. For neural forecasting models, data was organized into time-series objects by grouping over product and store identifiers. For scalability sake, numerical fields were downcast to lighter data types to optimize memory usage during model training on large datasets.

3 Overview of Forecasting Models and Implementation

This study evaluates performance results of a range of forecasting models commonly applied in time-series prediction. These models include tree-based ensembles, neural networks, and classical statistical approaches. Although the theoretical foundations of these models are well-established in the literature, our emphasis lies on their implementation, training configurations, and evaluation setup tailored to the demands of long-term brick-and-mortar retail forecasting.

All neural network models were implemented using the `Darts` library, which supports modular time-series architectures and standardized training workflows. Each model was trained across four experimental setups: (1) training on each product group independently using nonimputed data; (2) training on each product group using imputed data; (3) training across all product groups jointly on nonimputed data; and (4) training jointly using the imputed dataset. This design allowed us to assess the impacts of imputation and group-level modeling on sales forecasting results.

To optimize model performance, a hyperparameter search was conducted using the HEBO algorithm [33], a scalable Bayesian optimization method. Due to computational constraints, hyperparameter tuning was conducted on a randomly sampled 10% subset of nonimputed data points for ensemble models, and on a random 10% subset of time series with at least 730 observations for neural models. This subset was chosen to be representative of the overall data distribution, ensuring that optimized hyperparameters would generalize between training setups that are group-specific and full-category. The feature set used for all experiments was restricted to the subset identified as relevant through Boruta feature selection.

The naive mean model is used as a benchmark. It predicts the mean of each series for items present in the training set, and the group-level mean for items not seen during training.

By standardizing the implementation pipeline across models and varying experimental conditions, this setup provides a robust empirical basis for comparing forecasting approaches in a B&M retail setting.

3.1 Evaluation Metrics

In addition to standard evaluation metrics, we measured financial performance by computing demand error and bias [20] and WMAPE for both revenue and profit. True revenue was calculated as actual sales multiplied by the CPI-adjusted price, and true profit as actual sales multiplied by the unit margin (CPI-adjusted price minus cost of goods sold plus any rebate). Predicted revenue and profit were obtained by applying the same price and margin calculations to the model's sales forecasts. The WMAPE itself was computed as the sum of absolute errors divided by the sum of true values, yielding separate error rates for revenue and profit.

For the group-level evaluation, we aggregated both true and predicted revenue and profit by store zone and product group. For the series-level evaluation, we performed an analogous process at the individual product level. True and predicted revenue and profit were summed for each unique product identifier, WMAPE was computed for each series, and the final series-level performance was reported as the average of all per-series WMAPE values.

For experiments involving imputed data, evaluation was performed exclusively on the original (non-imputed) validation data points to ensure a fair comparison across models.

All experiments were executed on a workstation equipped with an AMD Ryzen Threadripper PRO 7985WX 64-core CPU, 256 GB of RAM, and 2xNVIDIA RTX 4080 GPUs.

4 Results

Model	RMSSE	MASE	MSE	RMSE	MAE	R2	ME	MFB	Theils Bias	Group Revenue WMAPE	Series Revenue WMAPE	Group Profit WMAPE	Series Profit WMAPE	Demand Error	Demand Bias
Case A: Individual Groups															
LGBM	**0.758**	0.847	**2.073**	**1.440**	0.605	0.507	-0.024	-0.036	0.000	**0.069**	0.231	0.117	0.252	0.670	-0.018
NH	0.930	1.205	3.203	1.790	0.882	0.209	0.001	0.002	0.000	0.192	0.337	0.201	0.343	0.840	0.017
NB	0.951	1.230	3.350	1.830	0.900	0.172	0.009	0.010	0.000	0.221	0.377	0.226	0.382	0.860	0.030
TFT	0.952	1.250	3.354	1.831	0.915	0.171	-0.013	-0.015	0.000	0.194	0.364	0.214	0.391	0.845	**-0.001**
XGB	0.759	**0.837**	2.076	1.441	**0.597**	0.506	-0.054	-0.081	0.001	0.072	**0.215**	0.096	0.231	0.670	-0.067
Case B: Whole Category															
LGBM	0.773	0.877	2.156	1.468	0.626	**0.511**	-0.070	-0.100	0.002	0.096	0.220	0.086	0.231	**0.665**	-0.087
XGB	0.784	0.879	2.215	1.488	0.628	0.498	-0.080	-0.114	0.003	0.112	0.218	0.082	**0.230**	0.675	-0.104
NH	1.463	1.648	7.716	2.778	1.176	-0.323	-0.062	-0.080	0.000	0.935	1.316	1.118	1.646	1.102	0.017
NB	1.482	1.649	7.920	2.814	1.177	-0.358	-0.073	-0.093	0.001	0.907	1.326	1.075	1.653	1.112	-0.006
TFT	1.319	1.493	6.274	2.505	1.066	-0.076	-0.101	-0.131	0.002	0.797	1.078	0.917	1.313	0.990	-0.073
Case C: Individual Groups, Imputed Train Data															
LGBM	1.205	1.504	5.341	2.311	1.095	-0.242	0.424	0.619	0.034	0.928	1.259	1.140	1.408	1.093	-0.294
NH	0.876	0.896	2.823	1.680	0.652	-0.018	-0.181	-0.330	0.012	0.286	0.452	0.150	0.410	0.981	-0.294
NB	1.083	1.084	4.316	2.077	0.789	-0.004	-0.153	-0.223	0.005	0.210	0.422	**0.040**	0.400	0.963	-0.214
TFT	1.076	1.253	4.264	2.065	0.912	0.008	**-0.000**	**-0.000**	0.000	0.153	0.456	0.248	0.544	0.936	0.042
XGB	0.831	0.878	2.540	1.594	0.639	0.409	-0.128	-0.187	0.006	0.219	0.386	0.166	0.378	0.746	-0.214
Case D: Whole Category, Imputed Train Data															
LGBM	1.010	1.144	3.679	1.918	0.816	0.144	-0.033	-0.048	0.000	0.311	0.685	0.322	0.788	0.878	-0.082
XGB	0.847	0.884	2.588	1.609	0.631	0.398	-0.197	-0.288	0.015	0.335	0.428	0.280	0.419	0.752	-0.304
NH	1.026	1.074	3.795	1.948	0.767	0.117	-0.172	-0.251	0.008	0.325	0.566	0.262	0.577	0.905	-0.266
NB	1.031	1.039	3.829	1.957	0.742	0.110	-0.218	-0.319	0.012	0.333	0.566	0.242	0.567	0.908	-0.330
TFT	0.992	1.114	3.549	1.884	0.795	0.174	-0.019	-0.028	0.000	0.252	0.421	0.259	0.474	0.856	-0.037

Table 3: Summary Statistics. Abbreviations: TFT: Temporal Fusion Transformer; NH: NHiTS; NB: NBEATS; LGBM: LightGBM; Best Results are Bolded and Second Best Underlined

We report both demand- and sales-based accuracy measures. The demand metrics follow standard forecasting practice, but our primary emphasis is on the sales metrics since accurate sales forecasts directly feed into the downstream pricing optimization.

In Table 3, the ensemble models, particularly LightGBM and XGBoost, exhibit strong performance in both individual-group raw data (Case A) and whole-category raw data (Case B) settings. When focusing on group-level evaluation metrics, specifically, group revenue and group profit WMAPE, which are critical for improved operational efficiency and financial results, ensemble methods demonstrate robust predictive capabilities. For example, in Case A, LightGBM achieves a group revenue WMAPE of 0.069 while XGBoost reports a competitive group profit WMAPE of 0.096. In the whole-category configuration (Case B), both LGBM and XGBoost maintain robust performance. Interestingly, an unexpected strong result is observed for NBEATS in the imputed individual-group setting (Case C), with a best-in-class group profit WMAPE of 0.040, though overall ensemble methods remain reliable across multiple metrics. These findings indicate that, depending on the metric prioritized by practitioners, ensemble models provide a reliable forecasting solution in both localized and aggregated settings.

For neural network-based models (including NBEATS, NHiTS, and the Temporal Fusion Transformer), the use of imputed data (Cases C and D) makes a significant difference. These models exhibit marked improvements, with reduced error rates compared to their nonimputed counterparts. This indicates that the imputation process, by restoring temporal continuity and mitigating the impact of missing values, is especially beneficial for complex architectures that are more sensitive to noise. Despite these gains, the best overall forecasting performance remains with the ensemble methods using the individual-group configuration.

An additional observation comes from the analysis of the demand bias metric. Although the bias is slightly negative for most models, including NBEATS (WC, non-imputed data), indicating a general tendency to underestimate actual demand, the aggregated predictions for the whole category in Figure 2 reveal a mild positive bias. This apparent discrepancy can be explained by the distribution of products: while the model tends to underestimate demand for

higher-priced items, it tends to overestimate for lower-priced items, which are more prevalent in the dataset. As a result, the overall effect skews the aggregate predictions toward a positive bias, despite localized underpredictions for premium products. Government interventions in pricing may further contribute to this imbalance by introducing systematic shifts in demand behavior and distorting historical patterns the model relies on.

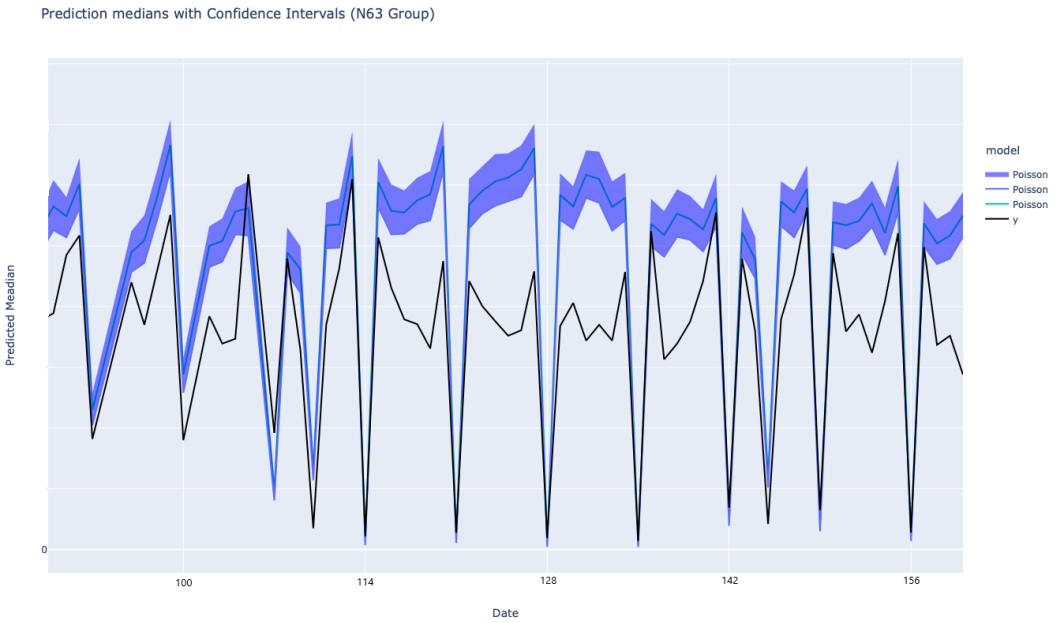
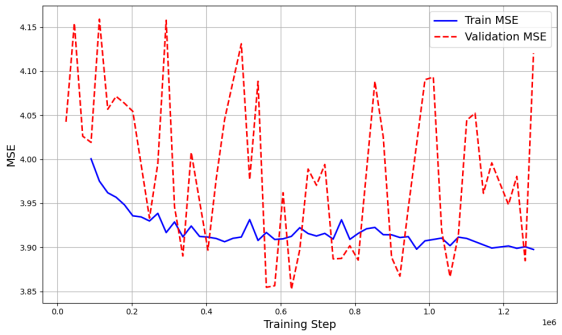
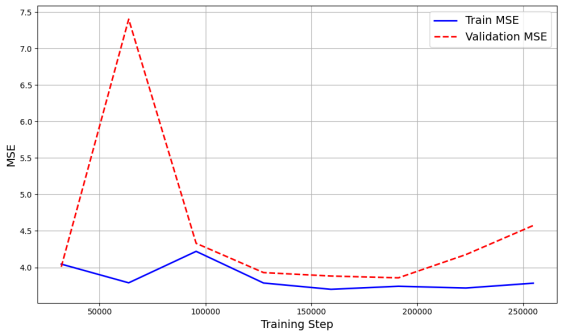


Figure 2: Prediction Medians with Confidence Intervals for NBEATS (WC, Non-Imputed Data) Model on N63 Group



(a) N-BEATS (WC, imp): Train vs. Validation MSE



(b) TFT (WC, raw): Train vs. Validation MSE

Figure 3: Comparison of Training and Validation MSE for N-BEATS and TFT models.

Beyond predictive accuracy, computational efficiency is another key factor used to evaluate forecasting models for practical use. Table 4 provides a detailed comparison of training times in the whole category setting for nonimputed and imputed data. The results make it evident that ensemble approaches, especially XGBoost, are highly efficient: they require considerably less memory and train much faster than neural network approaches. In contrast, neural models such as NBEATS and NHiTS demand significantly more resources, while the Temporal Fusion Transformer (TFT) incurs an even higher computational cost, particularly when trained on imputed data (which is expected since there are approximately ten times more training series for imputed data). For real-world applications where speed and accuracy matter, the lower computational footprint of the ensemble methods represents a clear advantage.

Model	Mean	Min	Max
NB (WC, imp)	171.555	171.555	171.555
NB (WC, raw)	15.818	15.818	15.818
NH (WC, imp)	114.340	114.340	114.340
NH (WC, raw)	10.000	9.986	10.014
TFT (WC, imp)	14678.233	14678.233	14678.233
TFT (WC, raw)	750.909	202.400	1299.418
LGBM (WC, imp)	35.566	31.047	42.010
LGBM (WC, raw)	11.237	7.383	15.533
XGBoost (WC, imp)	14.542	13.877	15.700
XGBoost (WC, raw)	6.191	5.486	8.041

Table 4: Training Time Statistics (in Minutes) Per epoch for neural networks, total for LGBM and XGB

5 Discussion

The preliminary findings presented in the Results section offer several important insights into retail sales forecasting using modern machine learning models. Notably, ensemble models such as LightGBM and XGBoost exhibit robust performance in both localized (individual-group) and aggregated (whole-category) configurations when using nonimputed data. These models consistently achieve lower error rates across multiple metrics, particularly in group revenue and group profit measures that are pivotal for pricing optimization, while also demonstrating a significant advantage in training time compared to more computationally intensive neural network approaches. This performance advantage aligns well with the broader literature on tabular data. For instance, Borisov et al. [14] highlight that despite the recent surge in deep learning innovations, gradient-boosted decision tree (GBDT) ensembles often remain the state-of-the-art for heterogeneous tabular data due to their robustness and lower sensitivity to data irregularities.

A key takeaway from our study is the importance of tailoring the modeling approach to the characteristics of the dataset. When data is segmented into individual groups and they do not share too much information between them, localized modeling can capture unique patterns more effectively. In contrast, training on the whole category does not introduce enough inter-group information to bring improvements, a phenomenon that resonates with the observations by McElfresh et al. [12], who note that differences in dataset properties, such as skewed feature distributions and irregularities, can diminish the relative advantage of complex neural network architectures over simpler, well-tuned tree-based methods.

Furthermore, experiments with imputed versus only processed data underscore the critical role of data quality and preprocessing. Deep neural networks, as discussed by Ramesh and Usman [13], tend to be more sensitive to inherent noise and missing values in tabular datasets. In our study, while imputation improved the performance of neural network models by restoring temporal continuity, ensemble methods on raw data still emerged as the best performers overall.

While recent work by Zalando [20] has demonstrated the presence of scaling laws for transformer-based models, where forecasting accuracy improves as training data size increases, such advantages were not observed in our experiments. This is likely due to the comparatively smaller scale of our dataset and the absence of centralized, high-density demand signals typical of e-commerce platforms. As a result, we did not observe a trade-off between computational cost and accuracy in favor of deep learning models. In our setting, ensemble methods remained both more efficient and more accurate, reinforcing their suitability for brick-and-mortar retail forecasting tasks characterized by fragmentation, intermittency, and limited historical coverage.

Practical Implications.
The demonstrated effectiveness of localized modeling using LightGBM or XGBoost indicates that segmenting retail data into individual product groups can lead to significant improvements in forecast accuracy. Retailers can leverage these insights to optimize inventory management, streamline supply chain operations, and design targeted marketing strategies. Additionally, the superior computational efficiency of tree-based methods makes them particularly well suited for real-time or near-real-time forecasting scenarios where fast turnaround is critical.

Limitations and Future Research.
Despite the promising results, several limitations of this study warrant discussion. One potential limitation is the reliance on aggregated error metrics, which, while informative, may obscure variability in model performance across different retail segments or temporal conditions. In addition, the experiments were conducted on datasets with specific characteristics, and the findings may not generalize to all types of retail environments, particularly those with extremely sparse or highly volatile data. Future research should investigate hybrid approaches that integrate the interpretability

and efficiency of tree-based models with the representation learning capabilities of neural networks. Additionally, further exploration into advanced data preprocessing techniques, such as more sophisticated imputation methods and feature transformation strategies, could help bridge the performance gap observed between nonimputed and imputed data experiments. Expanding the study to include a broader range of datasets and real-world retail scenarios would also provide deeper insights into the robustness and scalability of the proposed methods.

Overall, the findings suggest that for retail sales forecasting, leveraging a localized modeling strategy with robust, efficient algorithms such as LightGBM is advantageous. However, the debate between neural networks and boosted trees is not entirely one-sided. In scenarios where the dataset exhibits lower irregularity, neural networks might narrow the performance gap. These insights not only advance the state-of-the-art in retail sales forecasting but also provide a roadmap for future research and practical implementation in the retail sector.

6 Conclusion

This study underscores the critical importance of selecting a forecasting model that aligns with the unique challenges of brick-and-mortar retail environments. Our evaluation demonstrates that ensemble methods, particularly LightGBM and XGBoost, outperform complex neural network architectures in terms of both forecast accuracy and computational efficiency when applied to localized, individual-group data. The individual group approach enables models to more effectively adapt to unique group-specific dynamics, in contrast to combined data, where limited cross-group information sharing obscures critical nuances.

Although deep learning models have the potential to extract complex patterns from data, their sensitivity to data irregularities, such as missing values and noise, makes them less effective in environments characterized by high intermittency. Even with sophisticated imputation techniques like the SAITS model, the performance gains for neural networks remained limited. This can be attributed to the extensive amount of missing data in our setting, rather than a few isolated gaps, many series contain large contiguous sections of missing values. As a result, the imputed values introduce uncertainty that can overwhelm the underlying signal. Consequently, models trained on these imputed sequences may focus disproportionately on noisy reconstructions, further limiting their effectiveness compared to more robust ensemble methods.

The practical implications of our findings are significant. Retailers facing operational constraints can benefit from adopting a localized modeling strategy that leverages the strengths of tree-based ensembles, ensuring scalability, efficiency, and improved forecast accuracy. Moreover, the insights obtained from this work provide a valuable foundation for future research. Prospective studies should explore hybrid approaches that combine the robustness of tree-based models with the representational power of neural networks and investigate further enhancements in data preprocessing and imputation techniques.

In general, our work contributes to the evolving literature on forecasting in brick-and-mortar retail settings. It offers a systematic performance analysis of various forecasting models under realistic and dynamic conditions. In addition, it provides information for academics and practitioners to improve their accuracy of forecasts. Finally, it offers a robust foundation for broader advanced analytics workflows, enabling retail professionals to improve demand and pricing models, ultimately leading to more informed decision making and more beneficial business outcomes.

References

[1] George.E.P. Box and Gwilym M. Jenkins. *Time Series Analysis: Forecasting and Control*. Holden-Day, 1976.

[2] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 785–794. ACM, August 2016.

[3] 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. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[4] Boris N. Oreshkin, Dmitri Carpov, Nicolas Chapados, and Yoshua Bengio. N-beats: Neural basis expansion analysis for interpretable time series forecasting. In *International Conference on Learning Representations*, 2020.

[5] Cristian Challu, Kin G. Olivares, Boris N. Oreshkin, Federico Garza, Max Mergenthaler-Canseco, and Artur Dubrawski. N-hits: Neural hierarchical interpolation for time series forecasting, 2022.

[6] Bryan Lim, Sercan O. Arik, Nicolas Loeff, and Tomas Pfister. Temporal fusion transformers for interpretable multi-horizon time series forecasting, 2020.

[7] Yuqi Nie, Nam H Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *The Eleventh International Conference on Learning Representations*, 2023.

[8] Tianxiang Zhan, Yuanpeng He, Yong Deng, Zhen Li, Wenjie Du, and Qingsong Wen. Time evidence fusion network: Multi-source view in long-term time series forecasting, 2024.

[9] Tong Nie, Guoyang Qin, Wei Ma, Yuewen Mei, and Jian Sun. Imputeformer: Low rankness-induced transformers for generalizable spatiotemporal imputation. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, page 2260–2271, New York, NY, USA, 2024. Association for Computing Machinery.

[10] Luo donghao and wang xue. ModernTCN: A modern pure convolution structure for general time series analysis. In *The Twelfth International Conference on Learning Representations*, 2024.

[11] Zhijian Xu, Ailing Zeng, and Qiang Xu. FITS: Modeling time series with \$10k\$ parameters. In *The Twelfth International Conference on Learning Representations*, 2024.

[12] Duncan McElfresh, Sujay Khandagale, Jonathan Valverde, Vishak Prasad C, Benjamin Feuer, Chinmay Hegde, Ganesh Ramakrishnan, Micah Goldblum, and Colin White. When do neural nets outperform boosted trees on tabular data?, 2024.

[13] Abhinava Changa Ramesh and Muhammad Usman. *Empirical Study of Neural Network and other Machine Learning Models in Time Series Forecasting using Sales Data*. PhD thesis, Uppsala University, Disciplinary Domain of Science and Technology, Technology, Department of Civil and Industrial Engineering, Industrial Engineering and Management., 2024.

[14] Vadim Borisov, Tobias Leemann, Kathrin Seßler, Johannes Haug, Martin Pawelczyk, and Gjergji Kasneci. Deep neural networks and tabular data: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 35(6):7499–7519, 2024.

[15] Jintai Chen, Jiahuan Yan, Qiyuan Chen, Danny Ziyi Chen, Jian Wu, and Jimeng Sun. Excelformer: A neural network surpassing gbdts on tabular data, 2024.

[16] Alexander Holmberg and Dante Wesslund. Comparative analysis of gradient boosting decision trees and deep neural networks for tabular data, 2024.

[17] Myung-Jin Jun and. A comparison of a gradient boosting decision tree, random forests, and artificial neural networks to model urban land use changes: the case of the seoul metropolitan area. *International Journal of Geographical Information Science*, 35(11):2149–2167, 2021.

[18] Spyros Makridakis, Evangelos Spiliotis, and Vassilios Assimakopoulos. M5 accuracy competition: Results, findings, and conclusions. *International Journal of Forecasting*, 38(4):1346–1364, 2022. Special Issue: M5 competition.

[19] Tim Januschowski, Yuyang Wang, Kari Torkkola, Timo Erkkilä, Hilaf Hasson, and Jan Gasthaus. Forecasting with trees. *International Journal of Forecasting*, 38(4):1473–1481, 2022. Special Issue: M5 competition.